Approches statistiques pour l'analyse de données génomiques

Marie-Laure Martin-Magniette

Mémoire présenté en vue de l'obtention de
l'Habilitation à Diriger des Recherches

Spécialité : Mathématiques appliquées

Approches statistiques pour l'analyse de données génomiques

Présenté par

Marie-Laure Martin-Magniette

Soutenance le 4 décembre 2013 devant le jury composé de :

M. Christophe AMBROISE	Professeur à l'UEVE	Président du jury
M. Pierre BOUDINOT	Directeur de recherche à l'INRA	Rapporteur
M. David CAUSEUR	Professeur à l'AgroCampus Ouest	Rapporteur
Mme Christine DILLMANN	Professeur à l'Université Paris-Sud	Examinatrice
M. David HUNTER	Professeur à l'Université de Pennsylvanie	Rapporteur
M. Stéphane ROBIN	Directeur de recherche à l'INRA	Examinateur

Remerciements

Je remercie très chaleureusement David Causeur, David Hunter et Pierre Boudinot pour avoir accepté de rapporter mon mémoire, Christine Dillmann, Christophe Ambroise et Stéphane Robin pour avoir accepté d'être membre de mon jury. C'est un véritable honneur d'avoir pu vous réunir pour mon habilitation à diriger les recherches. Je vous suis extrêmement reconnais sante du temps que vous m'avez accordé et vous remercie sincèrement pour la discussion riche et très intéressante que nous avons eue le jour de la soutenance.

Souvent qualifiée "d'agrafe", de "charnière" ou encore de "cheville ouvrière", mes affiliations à l'Unité de Recherche en Génomique Végétale à Evry et l'Unité de Mathématiques et Informatique Appliquées 518 à Paris ont été possibles grâce à l'accompagnement permanent et constant de Jean-Jacques Daudin, Stéphane Robin, Michel Caboche et Alain Lecharny. Je vous remercie tout d'abord d'avoir créé ce poste, une première dans son genre, puis de m'avoir accordé de votre temps précieux à chaque fois que j'en ai eu besoin. Chacun à votre manière, vous m'avez aidée à trouver ma place à l'interface des statistiques et de la biologie moléculaire. Je tiens également à remercier Tristan Mary-Huard, mon cher co-bureau pendant 9 années, qui est peut-être le seul à réellement connaître les coulisses de mon aventure interdisciplinaire. Ce fut également un plaisir de travailler avec toi et j'espère que dans un futur pas trop éloigné nous arriverons à finaliser le travail sur le MFDR et que nos chemins se recroiseront pour mêler la génomique et la génétique dans un joli modèle statistique. Pour le moment, je te souhaite de prendre autant de plaisir que moi dans ton poste entre l'unité de Recherche en Génétique Végétale du Moulon et l'Unité de Mathématiques et Informatique Appliquées 518 à Paris.

Je remercie très sincèrement Pascal Massart et Sylvie Huet pour m'avoir fait découvrir et aimer le métier de chercheur. Depuis notre première rencontre, beaucoup de temps s'est écoulé et nos discussions s'espacent mais elles sont toujours aussi importantes pour moi. Votre bienveillance et votre disponibilité me sont précieuses. C'est une véritable chance de pouvoir profiter ainsi de votre expérience. Merci.

Les travaux présentés dans ce mémoire sont le résultat de travaux scientifiques réalisés avec de nombreux collaborateurs. Merci Julie et Jean-Jacques pour votre ténacité. Sans vous, de nombreux biais seraient restés dans les données issues des technologies à haut-débit. Merci Gilles et Cathy pour m'avoir initiée aux modèles de mélange. Je suis toujours surprise de voir tout le travail que nous avons réalisé sans vraiment avoir eu l'impression de travailler. Merci Andréa de nous avoir sollicités pour travailler sur les données de séquençage haut-débit et d'avoir réussi à intégrer le trio. Les projets ne manquent pas et ce sera toujours un grand plaisir de travailler avec vous. Merci Stéphane pour m'avoir fait redécouvrir les HMM. Ta générosité scientifique est exceptionnelle et je suis très heureuse de pouvoir en profiter. Ce fut un grand plaisir de co-encadrer avec toi les thèses de Caroline et Stevenn, que je remercie également. Ce n'est pas forcément facile d'être encadrante mais vous m'avez beaucoup facilité la tâche grâce à vos nombreuses qualités scientifiques et personnelles. Merci à toutes les personnes qui travaillent ou qui ont travaillé sur la plate-forme transcriptome de l'URGV pour avoir pris le temps de répondre à toutes mes questions parfois farfelues sur la génomique et les technologies. J'ai souvent mis votre patience à l'épreuve et vous m'avez souvent poussé dans mes derniers retranchements mais

je considère tout ce travail comme une belle réussite dont je suis très fière. Ensemble, nous avons réussi à construire et à faire évoluer une belle plate-forme, au service d'une communauté. Jean-Pierre, notre binôme sur la plate-forme a été riche tant sur le plan scientifique que personnel. Je te remercie pour ta ténacité et ton côté chercheur-aventurier qui auront eu pour conséquence de me faire aimer la génomique, au point de l'intégrer au cœur de mes projets de recherche. Merci Etienne pour avoir repris les flambeaux. C'est un réel plaisir de travailler avec toi et je suis convaincue que nous ne sommes qu'au début d'une belle et longue collaboration. Enfin, depuis peu je suis devenue responsable de l'équipe "bioinformatique pour la génomique prédictive" et je remercie Véronique, Cécile, Jean-Philippe, Guillem, Rim et Zakia de me faire confiance. C'est un vrai plaisir de travailler avec vous. J'aime votre vision de la bioinformatique mais aussi du travail et je suis sure que dans un futur relativement proche les gènes d'Arabidopsis auront de moins de moins de secrets pour nous. Merci également à Philippe pour l'installation et le maintien d'une informatique à la hauteur de nos exigences de bioinformaticiens et statisticiens. Sans serveurs et clusters, la génomique n'existerait pas.

Le travail ne constitue qu'une partie de ma vie et tout ce travail ne serait pas grand chose s'il n'était pas accompagné d'une vie personnelle tout aussi passionnante. Je remercie tout d'abord ma famille. J'aime votre joie de vivre, vos excentricités et votre bonne humeur. Merci à tous mes amis avec qui j'aime passer du temps et qui m'ont encouragée dans cette année 2013, peu ordinaire. Merci à mes deux pirates Paul et Adrien, qui rendent la vie tumultueuse et riante. Merci enfin à toi, Frédéric, pour tout ce que nous partageons ensemble depuis de nombreuses années et pour ton soutien avisé et constant dans tout ce que j'entreprends. La vie n'est pas un long fleuve tranquille et je suis heureuse de partager ta barque et peut-être bientôt ton bateau. Après une période pleine de tourbillons et de marmites, il me semble que nous voici enfin dans une accalmie dont je souhaite profiter pleinement en ta compagnie.

Contents

1 Introduction **7**

2 Model-based methods for clustering **9**
2.1 Variable selection in model-based clustering 9
2.1.1 Background 9
2.1.2 Variable role modeling 10
2.1.3 Theoretical properties 13
2.1.4 Procedure for performing the variable selection 15
2.1.5 Conclusion and perspectives 16
2.2 Model-based clustering under Markovian dependency 18
2.2.1 Background 18
2.2.2 Mixtures for estimating emission distributions 19
2.2.3 Variational Bayes approach for model aggregation 23
2.2.4 Conclusion and perspectives 26
2.3 Classification 27
2.3.1 Background 27
2.3.2 Conditional probabilities for a subset of observations in a HMM 28
2.3.3 Control of the false-positives in a mixture of two linear regressions 29
2.3.4 Conclusion and perspectives 30

3 Statistical approaches dedicated to genomic questions **31**
3.1 Context 31
3.2 Signal modeling for microarray data normalization 32
3.2.1 Background 32
3.2.2 Definition of an index to evaluate the gene-specific dye bias 33
3.2.3 Bleeding effects in multiple-target microarray 35
3.2.4 Generalization of the lowess procedure for multiple-target microarray . . . 37
3.3 Model-based clustering for whole-genome analysis 37
3.3.1 MixThres: Truncated Gaussian Mixtures for an hybridization threshold . 37
3.3.2 MultiChIPmix: Linear regression mixture for epigenomic data 40
3.3.3 TAHMMAnnot: Heterogeneous HMM for data of tiling array 44
3.4 Conclusions and perspectives 48

4 Contribution to genomic projects **51**
4.1 Creation of a genomic resource 51
4.1.1 Background 51
4.1.2 CATdb: a database for the transcriptomic platform of URGV 52
4.1.3 Statistics for the transcriptomic platform 54
4.1.4 Publications associated with the platform activity 57
4.2 Use of the transcriptomic resource to improve the structural annotation 58
4.2.1 Background 58
4.2.2 Characterization of novel genes 58

4.2.3 Conclusions 61
4.3 Other projects 62
4.3.1 Projects not dealing with omic data 62
4.3.2 Projects on control mechanisms of gene expression 62

5 Future projects on genomic networks **67**
5.1 Background 67
5.2 Open questions for improving the co-regulation networks 68
5.2.1 Modeling challenges 68
5.2.2 Model selection 69
5.2.3 Algorithms for variable selection 70
5.3 Inference of regulatory networks 70
5.3.1 Statistical learning and multivariate analysis for robust inference 70
5.3.2 High-dimensional regression for studying transcription factors 71

Chapter 1

Introduction

I began my scientific education at the university of Caen in 1993. After a licence in mathematics, I decided to pursue my formation by a master in mathematical ingeniery. During that year, I discovered statistics through both methodological concepts and various examples. At the end of this year, I was convinced that applied statistics could be an interesting domain. I spent one year at the university Paris-Sud and obtained a DEA on stochastic and statistical modeling. Then I started a PhD in 1998, co-supervised by Sylvie Huet (MIA-Jouy, INRA) and Pascal Massart (University of Paris-Sud, Orsay) and co-funded by INRA and the radioprotection division of EDF. I considered two statistical problems arising during the estimation of the hazard function of cancer death in Hiroshima. The first problem was dedicated to the estimation of the hazard function when the covariate is mismeasured. Application was done to estimate the cancer risk among the Hiroshima survivors taking into account that the received dose of radiation was mismeasured. The second problem was a proposal of new and more flexible modeling of the risk function in order to avoid the linear function usually used to link the risk function to the covariate.

From a statistical point of view, my PhD allowed me to acquire skills in survival models, models taking into account mismeasurement of the covariates and count process. From an application point of view, the analysis of the Hiroshima data was really an personal experience since I stayed three months at the Radiation Effect Research Fundation in Hiroshima to understand the specificity of the data and to develop the non-parametric estimation of the hazard function.

After my PhD, I obtained a post-doctoral position at INRA to work on multi-compartment models to propose dynamic modeling of epidemics. This project should provide a new insight on the Bovine spongiform encephalopathy. However despite my interest for the statistical modeling, I was disappointed by the absence of interactions between the different disciplinaries. Then, I accepted a position at the hospital of Nantes in the medical and public health information division, hosting biologists in human medicine and statisticians. Their mission is to develop databases for the hospital and to perform statistical studies on human health. In collaboration with Magali Giral and Yohann Foucher, I studied the impact of graft mass on the clinical outcome of kidney transplants. I really appreciated to work in closed collaboration with biologists. In retrospect, these four years have been rich in diverse and important experiences for the following.

In 2003, I succeeded to obtain a position as a researcher at INRA in the Plant Biology and Breeding department. The scientific positioning is original because I share my activity between the Unit of Plant Genomic (URGV) and the UMR AgroParisTech/INRA of Applied Mathematics and Informatics. At URGV I work with biologists to confront the statistical approaches to a biological reality. Moreover I have the chance to get involved in a lot of projects as soon as they start and it allows me to identify and solve their methodological needs. This position gave me the opportunity to develop methods and models for various kind of biological questions relying on the gene expression and their control mechanisms. Moreover, all these projects gave me opportunity of working with the most recent and sensitive biological technologies including all the different types of microarray and the high throughput sequencing. At UMR AgroParis-

Tech/INRA of Applied Mathematics and Informatics, I am in the team "Statistics and Genome" where members develop statistical models for analyzing molecular biology data. In this stimulating environment, I developed numerous collaborations in statistics around the problematics of URGV. It results in methodological developments in model-based clustering and signal modelings for normalization. From my point of view, this position is ideal because it allows me to stay in the core of the statistical research and at the same time to be deeply involved in the genomic field.

In the following manuscript, I focus on my research activity since 2003 and it is organized as follows: The first chapter is dedicated to methodological developments in model-based clustering, in particular the variable selection when observations are described by a large number of variables. Then considering that a spatial dependence exists between observations, I developed flexible models for estimating the emission distributions of an HMM. This first chapter finishes with some contributions on the observation classification based on the conditional probabilities. The second chapter is devoted to the developments of statistical methods particularly dedicated to the analysis of high throughput data coming from molecular biology. The third chapter is dedicated to biological projects. My main contribution is my involvement in the statistical analysis of the data produced by the transcriptomic platform of URGV. The fourth chapter is a presentation of the projects that I would like to develop in a near future.

Chapter 2

Model-based methods for clustering

2.1 Variable selection in model-based clustering

The works presented in this section correspond to results of the PhD of Cathy Maugis-Rabusseau, co-supervised with Gilles Celeux.

2.1.1 Background

The goal of clustering methods is to discover structures (clusters) among individuals described by several variables. Many clustering methods exist and roughly fall into two categories. The first one is based on similarity or dissimilarity measures. It gathers hierarchical clusterings, which build trees and also methods like K-means algorithm which classify data through a number of clusters fixed a priori. The second category is model-based methods which consist of using a model for clusters and optimizing the fit of the model and the data. In practice, each cluster is represented by a parametric distribution, like a Gaussian one and the entire dataset is modeled by a mixture of these distributions. An advantage of these model-based methods is to provide a statistically rigorous framework to assess the number of clusters and the role of each variable in the clustering process.

Cluster analysis is more and more concerned with large datasets where observations are described with many variables. In principle, the more information we have about each observation, the better a clustering method is expected to perform. However the structure of interest may often be contained in a subset of the available variables and a lot of variables may be useless or even harmful to detect a reasonable clustering structure. It is thus important to select the relevant variables from the cluster analysis view point. It is a recent research topic in contrast to variable selection in regression and classification models (Kohavi and John, 1997; Guyon and Elisseeff, 2003; Miller, 1990). This new interest for variable selection in clustering comes from the increasingly frequent use of these methods on high-dimensional datasets, such as transcriptome datasets. It is usually considered that co-expressed genes are often implicated in the same biological function and are potential candidates to be co-regulated genes (see for instance Sharan, Elkon, and Shamir, 2002, or Jiang, Tang, and Zhang, 2004, and references therein). Since the number of transcriptome experiments always increases, an experiment selection in the clustering procedure is desirable to reveal important biological phenomena.

Three types of approaches dealing with variable selection in clustering have been proposed. The first one includes clustering methods with weighted variables (see for instance Friedman and Meulman, 2004) and dimension reduction methods. For this latter, McLachlan *et al.* (2002) used a mixture of factor analyzers to reduce the extremely high dimensionality of a gene expression problem. A suitable Gaussian mixture family was considered in Bouveyron *et al.* (2007) to take into account the dimension reduction and the data clustering simultaneously. In contrast to this first method type, in the two last approaches the variable selection is done explicitly. In one hand, the so-called "filter" approaches select the variables before a clustering analysis (see for instance Dash *et al.*, 2002; Jouve and Nicoloyannis, 2005). Their main weakness is the

influence of selection step done independently of the clustering step. In the other hand, the so-called "wrapper" approaches combine variable selection and clustering. For distance-based methods, one can cite Fowlkes *et al.* (1988) for a forward selection approach with complete linkage hierarchical clustering, Devaney and Ram (1997) who proposed a stepwise algorithm where the quality of the variable subsets is measured with the COBWEB algorithm or the method of Brusco and Cradit (2001) based on the adjusted Rand index for K-means clustering. There also exists "wrapper" methods in the model-based clustering setting. When the number of variables is greater than the number of observations, Tadesse *et al.* (2005) proposed a fully Bayesian method using a reversible jump algorithm to simultaneously choose the number of mixture components and select variables. Kim *et al.* (2006) used a similar approach by formulating clustering in terms of Dirichlet process mixtures. In Gaussian mixture model clustering, Law *et al.* (2004) proposed to evaluate the importance of the variables in the clustering process via "feature saliencies" and used the *Minimum Message Length* criterion. Raftery and Dean (2006) recast the problem of comparing two nested variable subsets as a model comparison problem and addressed it using Bayes factor. An interesting aspect of their model formulation is that irrelevant variables are not required to be independent of the clustering variables. They avoid thus the unrealistic independence assumption between the relevant and irrelevant variables for the clustering, considered in Tadesse *et al.* (2005), Kim *et al.* (2006) and Law *et al.* (2004). In their model, the whole irrelevant variable subset depends on the whole relevant variables through a linear regression equation. However, some relevant variables are not necessarily required to explain all irrelevant variables in the linear regression and their introduction involves additional parameters without a significant increase of the log-likelihood.

During her PhD, Cathy Maugis-Rabusseau has improved their method by considering another type of relation between the irrelevant variables for clustering and the relevant ones. Two variable role modeling had been proposed, their theoretical properties were derived and the variable selection implementation was based on a backward stepwise algorithm. These are the works described below.

2.1.2 Variable role modeling

Model SR: first improvement

Consider n observations $\mathbf{y} = (\mathbf{y}_1', \ldots, \mathbf{y}_n')'$ described by Q variables ($\mathbf{y}_i'$ in $\mathbb{R}^Q$) and let define a latent variable Z_i which equals k if observation i arises from cluster k and 0 otherwise. The graph of the conditional dependence according to the partition of the Q variables proposed by Raftery and Dean (2006) is given in Figure 2.1. The variable Z indicates the ideal clustering, S denotes the set of relevant clustering variables and S^c, its complement in $\{1, \ldots, Q\}$, denotes the irrelevant variables. Nevertheless all relevant variables are not necessarily required to explain the irrelevant variables in the linear regression and their forced introduction involves to estimate additional parameters without a significant increase of the log-likelihood.

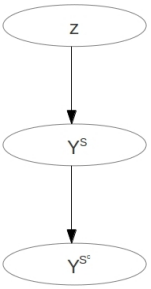

Figure 2.1: Conditional dependency graph associated to the variable repartition proposed by Raftery and Dean (2006)

In Maugis *et al.* (2009b), we improved their model by a model where irrelevant variables are

explained by a subset of relevant variables. This partition of the variable roles is illustrated on the graph of the conditional dependencies given Figure 2.2. The Z indicates the ideal clustering, the set of relevant clustering variables S is decomposed into R and $S \setminus R$. The variables R are those entering in the regression equation of the irrelevant variables S^c, the complement of S in $\{1, \ldots, Q\}$. Let $\mathcal{F}$ be the set of all variable subsets family, the variable role is thus summarized into a couple (S, R) belonging to

$$\mathcal{V} = \{(S, R); (S, R) \in \mathcal{F}^2,\ S \neq \emptyset,\ R \subseteq S\}.$$

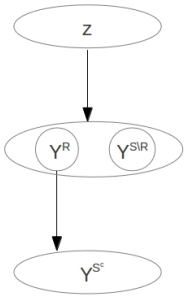

Figure 2.2: Conditional dependency graph associated to the variable repartition proposed by Maugis *et al.* (2009b).

The density family associated with a variable partition (S, R) was decomposed into two subfamilies of densities related to the variable partition :

1. On the relevant variable subset S, a Gaussian mixture was considered. It is characterized by its number of clusters K, varying from $K_{\min}$ to $K_{\max}$ defined by the user and its form m, essentially related to the assumptions on the component variance matrices. It means that

$$Z_i \sim \mathcal{M}(1, p_1, \ldots, p_K), \text{ with } \sum_{k=1}^{K} p_k = 1$$

and

$$(\mathbf{y}_i^S | Z_i = k) \sim \mathcal{N}(\mu_k, \Sigma_k),$$

with the variance matrix Σ_k satisfying the form m. Consequently, the likelihood associated with the relevant variable S for a given couple describing the mixture (K, m) is

$$f_{\text{clust}}(\mathbf{y}^S | K, m, \alpha) = \prod_{i=1}^{n} \sum_{k=1}^{K} p_k \Phi(\mathbf{y}_i^S | \mu_k, \Sigma_k) \tag{2.1}$$

where the parameter vector is $\alpha = (p_1, \ldots, p_{K-1}, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K)$.

2. The irrelevant variables of the subset S^c are explained by the variables of the subset R according to a multidimensional linear regression where the variance matrix are assumed to have a spherical, diagonal or general form. The likelihood associated with the linear regression of $\mathbf{y}^{S^c}$ on $\mathbf{y}^R$ is then

$$f_{\text{reg}}(\mathbf{y}^{S^c} | \mathbf{y}^R, r, a, \beta, \Omega) = \prod_{i=1}^{n} \Phi(\mathbf{y}_i^{S^c} | a + \mathbf{y}_i^R \beta, \Omega) \tag{2.2}$$

where a is the $1 \times \text{card}(S^c)$ intercept vector, β is the $\text{card}(R) \times \text{card}(S^c)$ coefficient regression matrix and Ω the $\text{card}(S^c) \times \text{card}(S^c)$ variance matrix.

The likelihood for a model is then given by

$$f(\mathbf{y}|K,m,r,S,R,\theta) = f_{\text{clust}}(\mathbf{y}^S|K,m,\alpha) f_{\text{reg}}(\mathbf{y}^{S^c}|\mathbf{y}^R,r,a,\beta,\Omega) \tag{2.3}$$

where the parameter vector $\theta = (\alpha, a, \beta, \Omega)$ belongs to a parameter vector set $\Upsilon_{(K,m,r,S,R)}$.

To fit well the data, we considered a collection of models $\mathcal{N}$ where the number of components of the mixture, the forms of the variance matrices as well as the variable partition vary. The models in competition are then compared with their integrated likelihoods, approximated by using the BIC approximation (Schwarz, 1978). The chosen model is defined as

$$(\hat{K},\hat{m},\hat{S},\hat{R}) = \underset{(K,m,S,R)\in\mathcal{N}}{\text{argmax}} \{\text{BIC}_{\text{clust}}(\mathbf{y}^S|K,m) + \text{BIC}_{\text{reg}}(\mathbf{y}^{S^c}|\mathbf{y}^R)\} \tag{2.4}$$

where $\text{BIC}_{\text{clust}}$ and BIC_{reg} are respectively the approximation of the integrated likelihood of the mixture and of the multivariate regression. The decomposition of the criterion as a sum of BIC criteria is a direct consequence of the conditional independences of the model given in (2.3).

Model SRUW: general variable role modeling

Thanks to several simulation studies, we realized that the model SR is not completely general since it does not allow some irrelevant variables to be independent and others to be dependent of the relevant variables simultaneously. In order to remedy to this drawback, in Maugis *et al.* (2009c), we proposed to refine the variable role modeling to take into account the possibility that some irrelevant clustering variables could be independent of all the relevant clustering variables and others could be linked to some relevant variables at the same time. This new variable partition is summarized in Figure 2.3 and such modeling allows us to define completely the variable role by subsetting the relevant variables for the clustering S, the redundant variables, U defined as irrelevant variables linked to some relevant variables, R and, the independent variables W defined as irrelevant variables independent of all the relevant variables.

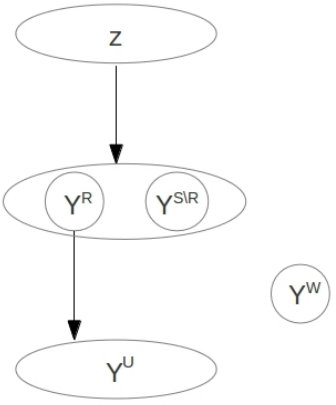

Figure 2.3: Conditional dependency graph associated to a variable partition $\mathbf{V} = (S,R,U,W)$.

The variable role is thus summarized into a couple (S,R,U,W) belonging to

$$\mathcal{V} = \left\{ (S,R,U,W) \in \mathcal{F}^4;\ \begin{array}{l} S \oplus U \oplus W = \{1,\ldots,Q\} \\ S \neq \emptyset, R \subseteq S \\ R = \emptyset \text{ if } U = \emptyset \text{ and } R \neq \emptyset \text{ otherwise} \end{array} \right\}.$$

The density family associated with a variable partition (S,R,U,W) is decomposed into three subfamilies of densities related to the three possible variable roles.

1. On the relevant variable subset S, a Gaussian mixture is considered as in (2.1).

2. The variables of the subset U are explained by the variables of the subset R according to a multidimensional linear regression where the variance matrix could have a spherical, diagonal or general form. The likelihood associated with the linear regression of $\mathbf{y}^U$ on $\mathbf{y}^R$ is defined as in (2.2):

$$f_{\text{reg}}(\mathbf{y}^U|\mathbf{y}^R, r, a, \beta, \Omega) = \prod_{i=1}^{n} \Phi(\mathbf{y}_i^U|a + \mathbf{y}_i^R\beta, \Omega)$$

where a is the $1 \times \text{card}(U)$ intercept vector, β is the $\text{card}(R) \times \text{card}(U)$ coefficient regression matrix and Ω is the $\text{card}(U) \times \text{card}(U)$ variance matrix.

3. The marginal distribution of the data on the variable subset W, which contains the variables independent of all relevant variables, is assumed to be a Gaussian distribution with mean vector γ and variance matrix τ. The form of the variance matrix τ could be spherical or diagonal. The associated likelihood on W is then

$$f_{\text{indep}}(\mathbf{y}^W|l, \gamma, \tau) = \prod_{i=1}^{n} \Phi(\mathbf{y}_i^W|\gamma, \tau).$$

The likelihood for a model $(K, m, r, l, \mathbf{V})$ is given by

$$f(\mathbf{y}|K, m, r, l, \mathbf{V}, \theta) = f_{\text{clust}}(\mathbf{y}^S|K, m, \alpha) f_{\text{reg}}(\mathbf{y}^U|\mathbf{y}^R, r, a, \beta, \Omega) f_{\text{indep}}(\mathbf{y}^W|l, \gamma, \tau) \tag{2.5}$$

where θ denotes the parameter vector $(\alpha, a, \beta, \Omega, \gamma, \tau)$. This model, named SRUW, is a generalization of the model proposed in Maugis *et al.* (2009b), since this latter can be interpreted as a SRUW model with $U = S^c$ and $W = \emptyset$. As for the model SR, approximation of the integrated log-likelihood derived from (2.5) leads that the selected the model is the one maximizing:

$$\text{BIC}_{\text{clust}}(\mathbf{y}^S|K, m) + \text{BIC}_{\text{reg}}(\mathbf{y}^U|r, \mathbf{y}^R) + \text{BIC}_{\text{indep}}(\mathbf{y}^W|l), \tag{2.6}$$

where the three functions are the associated BIC approximations of f_{clust}, f_{reg} and f_{indep}. As for model SR, the decomposition as the sum of three terms is a direct consequence of the conditional independences of the model SRUW given in (2.5)

2.1.3 Theoretical properties

For both models, the theoretical properties are established. Necessary and sufficient conditions are given to ensure the identifiability of the model collections and a consistency theorem of the variable selection criterions are stated. In practice, the theoretical properties for the SR model were generalized to the SRUW model. However this latter being the most general, I give here its theoretical properties.

Theorem 2.1.1 *Let $\Theta_{(K,m,r,l,\mathbf{V})}$ be a subset of the parameter set $\Upsilon_{(K,m,r,l,\mathbf{V})}$, such that elements $\theta = (\alpha, a, \beta, \Omega, \gamma, \tau)$*

- *contain distinct couples (μ_k, Σ_k) fulfilling $\forall s \subsetneq S, \exists (k, k'), 1 \leq k < k' \leq K$*

$$\mu_{k,\bar{s}|s} \neq \mu_{k',\bar{s}|s} \text{ or } \Sigma_{k,\bar{s}|s} \neq \Sigma_{k',\bar{s}|s} \text{ or } \Sigma_{k,\bar{s}\bar{s}|s} \neq \Sigma_{k',\bar{s}\bar{s}|s}, \tag{2.7}$$

  where $\bar{s}$ denotes the complement in S of any nonempty subset s of S

- *if $U \neq \emptyset$,*
  - *for all variables j of R, there exists a variable u of U such that the restriction β_{uj} of the regression coefficient matrix β associated with j and u is not equal to zero.*
  - *for all variables u of U, there exists a variable j of R such that $\beta_{uj} \neq 0$.*

- *Parameters Ω and τ strictly respect the forms r and l respectively: They are both diagonal matrices with at least two different eigenvalues if r and l are diagonal matrices and Ω has at least a non-zero entry outside the main diagonal if r is a general matrix.*

Let $(K, m, r, l, \mathbf{V})$ and $(K^\star, m^\star, r^\star, l^\star, \mathbf{V}^\star)$ be two models. If there exist $\theta \in \Theta_{(K,m,r,l,\mathbf{V})}$ and $\theta^\star \in \Theta_{(K^\star,m^\star,r^\star,l^\star,\mathbf{V}^\star)}$ such that

$$f(.|K, m, r, l, \mathbf{V}, \theta) = f(.|K^\star, m^\star, r^\star, l^\star, \mathbf{V}^\star, \theta^\star)$$

then $(K, m, r, l, \mathbf{V}) = (K^\star, m^\star, r^\star, l^\star, \mathbf{V}^\star)$ and $\theta = \theta^\star$ (up to a permutation of mixture components).

The conditions to ensure identifiability sound rather technical, but appeared to be natural. Condition (2.7) is the main condition : If the parameters $\mu_{k,\bar{s}|s}$, $\Sigma_{k,\bar{s}|s}$ and $\Sigma_{k,\bar{s}\bar{s}|s}$ are identical for all clusters then the identifiability could not be ensured because the regression density of $\bar{s}$ on s could be factorized from the Gaussian mixture and regrouped with the regression density of S^c on R. The second condition ensures that the subsets R and U are well defined. The regression model proposed to link the redundant variables to a subset of relevant variables prohibits to exchange a redundant variable and a relevant variable. The third condition speaks by itself (for instance a diagonal variance matrix cannot be spherical). From a practical point of view, those conditions appear to be quite reasonable and are expected to hold in real situations. It is not possible to formally check identifiability, but situations where the main Condition (2.7) would not be fulfilled are exotic.

The second theorem stated the consistency property for the SRUW model. We proved that the probability of selecting the true variable partition $\mathbf{V}_0 = (S_0, R_0, U_0, W_0)$ by maximizing criterion (2.6) approaches 1 as $n \to \infty$ when the sampling distribution is one of the densities in competition and the true model (K_0, m_0, r_0, l_0) is known. Details are available in Maugis *et al.* (2008).

Theorem 2.1.2 *Denoting h the density function of the sample $\mathbf{y}$ and*

$$\hat{\theta} = \underset{\theta \in \Theta_{(K,m,r,l,\mathbf{V})}}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} \ln\{f(\mathbf{y}_i|K, m, r, l, \mathbf{V}, \theta)\}$$

assuming that

(H1) The density h is one of the densities in competition: There exists a unique known model $(K_0, m_0, r_0, l_0, \mathbf{V}_0)$ such that $h = f(.|\theta^\star_{(K_0, m_0, r_0, l_0, \mathbf{V}_0)})$ where

$$\theta^\star = \underset{\theta \in \Theta_{(K,m,r,l,\mathbf{V})}}{\operatorname{argmin}} KL[h, f(.|K, m, r, l, \mathbf{V}, \theta)]$$

with $KL[h, f] = \int \ln\left\{\frac{h(x)}{f(x)}\right\} h(x)dx$ is the Kullback-Leibler divergence between the densities h and f.

(H2) The vectors $\theta^\star$ and $\hat{\theta}$ are supposed to belong to a compact subspace $\Theta'_{(K_0,m_0,r_0,l_0,\mathbf{V})}$.

The variable partition $\hat{\mathbf{V}} = (\hat{S}, \hat{R}, \hat{U}, \hat{W})$ maximizing Criterion (2.6) with fixed (K_0, m_0, r_0, l_0) is such that

$$P(\hat{\mathbf{V}} = \mathbf{V}_0) = P((\hat{S}, \hat{R}, \hat{U}, \hat{W}) = (S_0, R_0, U_0, W_0)) \underset{n\to\infty}{\to} 1.$$

for both theorems, proofs of these results are detailed in Maugis *et al.* (2008).

2.1.4 Procedure for performing the variable selection

An exhaustive search of the optimal model is impossible in most situations. Consequently, we proposed a two-nested-step algorithm. The first step consists in defining the best variable partition for a given mixture model characterized by its number of component K and the form of the variance matrices m. It is done by comparing the integrated likelihoods. The second step consists in determining the best model of the collection by comparing the integrated likelihoods when the mixture varies. First the idea of the algorithm is described for the model SR and then its generalization for the model SRUW is presented.

For the model SR

The most delicate step of the procedure is to determine the variable partition for a given mixture defined by the couple (K, m). In Maugis *et al.* (2009b), an embedding backward stepwise procedure is proposed: All the variables $\{1, ..., Q\}$ are declared relevant at the beginning and then exclusion steps and inclusion steps are alternated to remove or include one-by-one some of them until that the relevant variable set becomes stable. It allows one to define the relevant clustering variable subset $\hat{S}(K, m)$. the subset $\hat{R}(K, m)$ is then obtained using a backward stepwise algorithm for the regression of $\mathbf{y}^{S^c}$ on $\mathbf{y}^S$. In this latter algorithm, all the variables $\{1, ..., \hat{S}(K, m)\}$ are considered to explain $\mathbf{y}^{S^c}$ and then exclusion and inclusion steps are alternated to remove or include one-by-one some of them until that the relevant variable set explaining $\mathbf{y}^{S^c}$ becomes stable.

The idea of the algorithm for defining $\hat{S}(K, m)$ is described below. At each step of the algorithm the variable $\{1, \ldots, Q\}$ is divided into three subgroups: S the set of selected clustering variables, j the candidate variable being considered for inclusion into or exclusion from the set of clustering variables and U the irrelevant variable set. The decision of exclusion (resp. inclusion) of variable j from (resp. in) the set of clustering variables is made by the comparison of two models:

- Model $M_1(K, m)$ specifies that given $\mathbf{y}^S$, $\mathbf{y}^j$ does not provide additional information for the clustering and is explained by a subset $\mathbf{y}^{R[j]}$ of $\mathbf{y}^S$.
- Model $M_2(K, m)$ specifies that given $\mathbf{y}^S$, $\mathbf{y}^j$ provides additional information for the clustering.

The two models are compared with the Bayes factor approximated by:

$$\mathrm{BIC}_{\mathrm{diff}}(\mathbf{y}^j|K, m) = \mathrm{BIC}_{\mathrm{clust}}(\mathbf{y}^S, \mathbf{y}^j|K, m) - \left\{\mathrm{BIC}_{\mathrm{clust}}(\mathbf{y}^S|K, m) + \mathrm{BIC}_{\mathrm{reg}}(\mathbf{y}^j|\mathbf{y}^{R[j]})\right\}. \quad (2.8)$$

If $\mathrm{BIC}_{\mathrm{diff}}(\mathbf{y}^j|K, m)$ is positive, then $M_2(K, m)$ is chosen otherwise, $M_1(K, m)$ is chosen. To choose the candidate variable in an exclusion step, all variables of S are considered and j is defined as the one minimizing $\mathrm{BIC}_{\mathrm{diff}}(\mathbf{y}^j|K, m)$. In an inclusion step, all variables of S^c are considered and j is the one maximizing $\mathrm{BIC}_{\mathrm{diff}}(\mathbf{y}^j|K, m)$.

The backward stepwise algorithm for the regression is similar and is described in Maugis *et al.* (2009b).

Adaptation for the model SRUW

The algorithm is analogous to the one previously described for the model SR. The first step consists in defining $\hat{S}(K, m)$ using a backward stepwise algorithm. The main change comes from the three situations that can occur for the candidate variable j:

- $M1(K, m)$: Given $\mathbf{y}^S$, $\mathbf{y}^j$ provides additional information for the clustering,
- $M2(K, m)$: Given $\mathbf{y}^S$, $\mathbf{y}^j$ does not provide additional information for the clustering but has a linear link with the variables of $R[j]$ (the nonempty subset of S containing the relevant variables for the regression of $\mathbf{y}^j$ on $\mathbf{y}^S$),

- $M3(K,m)$: Given $\mathbf{y}^S$, $\mathbf{y}^j$ is independent of all the variables of S,

The comparison of the three models is less straightforward than a comparison of two models as previously. Fortunately by pointing out that the density under $M3(K,m)$ can be written as the density under $M2(K,m)$ with an empty subset $R[j]$, it allows us to recast the comparison of the three models into the comparison of two models and to use the criterion (2.8) constructed for the SR model.

To summarize, the algorithm first consists of separating variables into relevant and irrelevant variables. Second, the irrelevant variables are partitioned into redundant and independent variables according to the conditioning relevant variables required in the regression. It remains then to determine the set of conditioning relevant variables for the multidimensional regression of the redundant variables and the general variance structures.

1. For each mixture model (K,m):
   - Determine the variable partition into $\hat{S}(K,m)$ and $\hat{S}^c(K,m)$ by the stepwise selection algorithm.
   - Divide the variable subset $\hat{S}^c(K,m)$ into $\hat{U}(K,m)$ and $\hat{W}(K,m)$ by the stepwise regression algorithm. For each variable j belonging to $\hat{S}^c(K,m)$, if the variable subset $R[j]$ of $\hat{S}(K,m)$ explaining j by a linear regression is empty, $j \in \hat{W}(K,m)$, otherwise, $j \in \hat{U}(K,m)$.
   - For each form r:
     * Determine the variable subset $\hat{R}(K,m,r)$, included into $\hat{S}(K,m)$ and explaining the variables of $\hat{U}(K,m)$ by using a stepwise regression algorithm with the fixed form regression model r.
     * For each form l, compute $\hat{\theta}$ and the criterion (2.6) for $\hat{S}(K,m)$,$\hat{R}(K,m,r)$,$\hat{U}(K,m)$,$\hat{W}(K,m)$.
2. Then determine the best model of the collection which maximizes (2.6).

Thanks to the trick and despite the three possible roles of each variables, the complexity of the algorithm is not increased compared with the algorithm of model SR.

2.1.5 Conclusion and perspectives

Variable selection in clustering is a recent research topic in contrast to variable selection in regression and classification models and this new interest comes from the increasingly frequent use of these methods on high-dimensional datasets. Our works led to a new modeling principle of the variable role in a model-based clustering setting. The SRUW model is the most versatile since it recovers all the possible variable roles: Significant (S), redundant (U in relation with R) and noisy (W). Moreover all previously studied variable selection models in model-based clustering can be obtained as particular SRUW models.

From our experience on several datasets, variable selection in a model-based clustering setting improves the clustering and its interpretation. It can happen that the data clustering is modified. In such a case, the clusters are more homogeneous. In the other case, which can often happen, the data clustering is not modified but the variable selection model provides an useful interpretation of the variable role. In Maugis *et al.* (2009a) we illustrated the use of SRUW model for discovering the function of orphan genes of *Arabidopsis thaliana* plant from a transcriptome dataset. We also compared our approach to sparse k-means and concluded that it is competitive (Celeux *et al.*, 2011)

Because in real dataset, missing values occur, we also studied this aspect for data missing at random in Maugis-Rabusseau *et al.* (2012). A common practice is to remove observations with missing values or to use imputation values to replace missing values. However it is known to have an important impact on the clustering result (Celton *et al.*, 2010). Consequently, we extended the two models SR and SRUW with a management of the missing values inside the

statistical model. It avoids a pre-processing step, often tricky and performs the variable selection and the clustering of data with missing values. It required a new strategy to calculate the model selection criterion via the explicit expression of the observed log-likelihood and a new parameter estimation method (see Sections 3.2.1 and 3.2.2 of Maugis-Rabusseau *et al.*, 2012). Applications on simulated data and real data showed that the variable selection procedure combined with an imputation method could have difficulties to find the variable partition. Whereas our method seems to be more reliable than imputation methods. It was able to find the true model up to 20% of missing values, the error rates were among the smallest.

After studying the variable selection problem in model-based clustering, we focused on generative models, as Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA), which estimate the class-conditional densities. Generative models are less parsimonious than predictive and non probabilistic methods and those last methods are generally preferred to generative models when the number of predictors is large in regard to the number of objects in the training set. However, generative models have some advantages since they allow us to determine the marginal density of the data. As noted in Hastie *et al.* (2009), LDA and QDA are widely used and perform well on an amazingly large and diverse set of classification problems. Moreover LDA is regarded as a reference method by many practitioners, and an advantage of LDA over QDA is that it is a more parsimonious method. There exists quite efficient methods to select predictors in the LDA context. Efficient stepwise variable selection procedures are available in most statistical softwares (see McLachlan, 1992, Section 12.3.3). On the contrary, there is less available material for QDA (Young and Odell, 1986), and as far as we know, no variable selection procedure for QDA is available in standard statistical softwares. However in the last few years, there is a renewal of interest in this topic. Zhang and Wang (2008) proposed a variable selection procedure for QDA based on a BIC criterion and Murphy *et al.* (2010) adapted the variable selection procedure of Raftery and Dean (2006) to the supervised classification context. In Maugis *et al.* (2011) we considered the general variable role modeling proposed in the model SRUW in model-based discriminant analysis. Numerical experiments highlighted the potentially great interest of our variable selection procedure to improve the classification performances of non linear Gaussian classification models. Actually those models involve many parameters when the number of variables is large with respect to the training sample size. But our variable selection procedure allows us to overcome the dimensionality problem leading to powerful classifiers with a nice interpretation of variable roles. Our opinion is that our methodology is able to make the non linear generative classification methods such as quadratic discriminant analysis much more efficient in high dimensional contexts and competitive with gold standard classifiers such as LDA, logistic regression, k-nearest neighbor classifier or support vector classifiers in many situations.

These modelings are interesting only if they are available. For this reason, a huge work of programming has been done to propose softwares available at `http://www.math.univ-toulouse.fr/~maugis/`. They are based on the Mixmod software (Biernacki *et al.*, 2006). Nevertheless, algorithms are greedy in computation time and could discourage the users. To tackle this problem, it seems that the key is to change the initialization of the stepwise procedure, because it remains a delicate step and the main drawback of our models. Currently the initialization of the relevant set of variables is either all variables (backward procedure) or none (forward procedure). Nevertheless, these two procedures are the extremes and the best solution is probably between these two strategies, by defining a subset of relevant variables. First idea is to use prior information on the variables but it is rarely available. Consequently it is better not to rely on and to develop new approaches to improving the initialization step.

The idea of these developments comes from the co-expression study of transcriptome data. Since recently, transcriptome data are produced with microarray and observations are hybridization signals that it is possible to model with Gaussian distributions. In recent years, gene expression studies have increasingly made use of High Throughput Sequencing (HTS) technology which provides counts per gene. In this new context, it again requires to think about the definition of

a new distribution of the component and a new modeling of the relationship between relevant and irrelevant variables. So many open questions remain because only Gaussian mixtures are for the moment considered.

2.2 Model-based clustering under Markovian dependency

The works presented in this section correspond to the results of the PhD of Caroline Bérard and of Stevenn Volant, co-supervised with Stéphane Robin.

2.2.1 Background

Hidden Markov models (HMM) constitute an efficient technique of unsupervised classification for longitudinal data. In such models, the neighborhood structure is accounted for via a Markov dependency between the latent variables, whereas the conditional distribution of the observation is ruled by the so-called 'emission' distribution. This latter being usually a Gaussian distribution, HMM can be viewed as a generalization of the classical mixture model. HMM were applied in many fields including signal processing (Rabiner, 1989), epidemiology (Sun and Cai, 2009) or genomics (Li *et al.*, 2005; Durbin *et al.*, 1998). In most cases, the emission distributions are given a specific form from a parametric class such as Gaussian, Gamma or Poisson. This may lead to a poor fit, when the distribution of the data is far from the chosen class. Efforts were made to propose more complex models capable of fitting skewed or heavy-tailed distribution, such as the multivariate normal inverse Gaussian distribution proposed in Chatzis (2010). Nevertheless when a priori knowledge on the clusters or on their numbers is available, the emission distributions are often multimodal and the proposed models are not adapted for these cases. I faced this situation when the high-throughput technologies in molecular biology became very perfected. An example concerns the comparison of two transcriptomes when data are generated with a microarray where probes cover the whole genome and where genes are observed through several adjacent probes. From a biological point of view, four probe groups are expected and the conditional distributions of the observations are usually multimodal. This question is considered in Section 3.3.3

In the framework of the model-based clustering where no spatial dependence of the latent variable is taken into account, the emission distribution modeling has been more investigated (Hennig, 2010). Most methods start with a large number of Gaussian distributions and finish with a smaller one. At each step two Gaussian components are merged into one by re-estimating the parameters. It leads to better separate the clusters, but their distribution are still assumed to be unimodal. A great interest has been recently paid to the definition of more flexible models using mixture as emission distributions (Li, 2005; Baudry *et al.*, 2010). These methods start with a large number of Gaussian distributions and aim at combining them, rather than at merging them. This idea can be viewed as semi-parametric approach as the shape of the distribution of each component of these mixtures is hoped to have a weak influence on the estimation of the emission distributions. The main difficulty is to combine the components and classical clustering algorithms are generally used: K-means approach (Li, 2005) or hierarchical clustering (Baudry *et al.*, 2010).

In Volant *et al.* (2013) we extended this semi-parametric modeling to the HMM context. We first showed that the inference can be fully achieved using the EM algorithm of Dempster *et al.* (1977) and proposed a hierarchical method for the initialization. To estimate the number of hidden states, BIC-like criteria were derived and had been compared in a simulation study to provide an effective procedure of the emission distributions of a general HMM. This work is described below.

In Volant *et al.* (2012) we focused on a two-state HMM, where the emission distribution of the first state is completely known whereas the second emission distribution is unknown and can be multimodal. To retrieve the binary latent variables $\{Z_t\}$ associated with each observation, we considered a finite collection of mixtures with a varying number of components and gathered information provided by each of them using Bayesian model averaging (BMA). We proposed

variational weights and compared them in a simulation study. This work is described in Section 2.2.3

2.2.2 Mixtures for estimating emission distributions

Emission distribution modeling

Consider n observations $y_1, \ldots, y_n$, each one described by a reasonable number of variables, denoted Q. The latent variable $\{Z_t\}$ is a D-state homogeneous Markov chain with transition matrix $\Pi = [\pi_{dd'}]$ and stationary distribution q. The observations $\{Y_t\}$ are independent conditionally to the hidden state with emission distribution ψ_d $(d = 1, ..., D)$. Figure 2.4 is a representation of the conditional dependencies of a classical HMM.

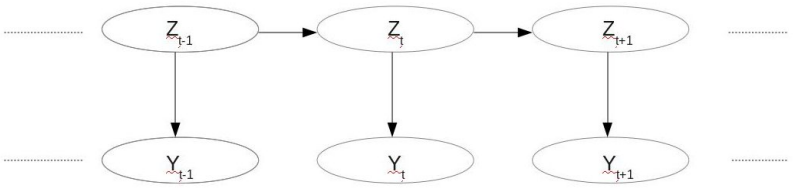

Figure 2.4: Conditional dependency graph in a HMM

In most cases, the emission distributions are given a specific form from a parametric class such as Gaussian, Gamma or Poisson. This may lead to a poor fit, when the distribution of the data is far from the chosen class, especially when the emission distributions are multimodal. To tackle this problem, we proposed to model each emission distribution ψ_d as a mixture of K_d parametric distributions:

$$(Y_t|Z_t = d) \sim \psi_d = \sum_{k=1}^{K_d} \lambda_{dk}\phi(.; \gamma_{dk}), \tag{2.9}$$

where λ_{dk} is the mixing proportion of the k-th component from cluster d $(\forall k \in \{1, ..., K_d\}$, $0 < \lambda_{dk} < 1$ and $\sum_k \lambda_{dk} = 1)$ and $\phi(.; \gamma)$ denotes a parametric distribution known up to the parameter vector γ. Since the distributions ψ_d are themselves mixtures, a second hidden variable $\{W_t\}$ is introduced to indicate the component k within state d, denoted (dk). It is itself a Markov chain with the transition matrix $\Omega = \left[\omega_{(dk),(d'k')}\right]$, where

$$\omega_{(dk),(d'k')} = \pi_{dd'}\lambda_{d'k'}, \tag{2.10}$$

so the transition between (dk) and $(d'k')$ only depend on the hidden state Z at the previous time. In this model, the graph of the conditional dependencies includes the new latent variables as illustrated in Figure 2.5. Introduction of the second latent variables $\{W_t\}$ is instrumental in developing the estimation procedure. Nevertheless the role of the $\{W_t\}$ is limited to a technical role because they will never used for the observation classification.

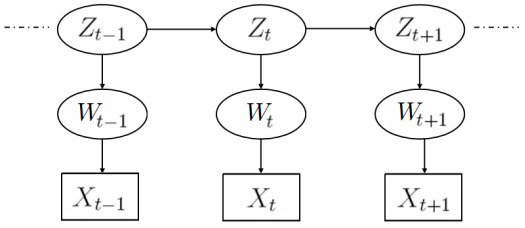

Figure 2.5: Conditional dependency graph in an HMM with mixtures as emission distributions

Finally to fit well the emission distributions, we considered a collection of models $\mathcal{M}$ where each model m is defined by D and the D-uplet $(K_1, \ldots, K_D)$ specifying the number of hidden states and the number of components within each hidden state.

$$\mathcal{M} = \left\{ m := (D, K_1, \ldots, K_D); 1 \leq D \ \ \forall d \ \ K_d \geq 1 \text{ with } \sum_{d=1}^{D} K_d = K \right\},$$

The choice of K is left to the user because as the $\{W_t\}$, it only plays a subsidiary role for the classification problem.

Inference for a given model $m \in \mathcal{M}$

The most common strategy for maximum likelihood inference of HMM relies on the EM algorithm (Dempster *et al.*, 1977; Cappé *et al.*, 2005). Despite the existence of two hidden layers Z and W, this algorithm could be applied by using the following decomposition of the log-likelihood

$$\log P(Y) = \mathbb{E}_Y[\log P(Y, Z, W)] - \mathbb{E}_Y[\log P(Z, W|Y)]$$

where $\mathbb{E}_Y$ stands for the conditional expectation, given the observed data Y. The E-step consists in the calculation of the conditional distribution $P(Z, W|Y)$ using the current value of the parameter θ^h. As $P(Z, W|Y) = P(Z|Y)P(W|Z, Y)$, the conditional distribution of the latent variables $P(Z, W|Y)$ can be calculated in two steps. First, $P(Z_t = d|Y) := \tau_{td}$ is calculated via the forward-backward algorithm (Rabiner, 1989) which only necessitates the current estimate of the transition matrix Π^h and the current estimates of the emission densities at each observation point: $\psi_d^h(Y_t)$. Second, $P(W_t = dk|Z_t = d, Y) := \delta_{tdk}$ is given by

$$\delta_{tdk}^{(h)} = \frac{\lambda_{dk}^{(h)} \phi(Y_t, \gamma_{dk}^{(h)})}{\sum_{j=1}^{K_d} \lambda_{dj}^{(h)} \phi(Y_t, \gamma_{dj}^{(h)})}.$$

For the M-step, the completed log-likelihood is developed as

$$\mathbb{E}_Y[\log P(Y, Z, W)] = \mathbb{E}_Y[\log P(Z)] + \mathbb{E}_Y[\log P(W|Z)] + \mathbb{E}_Y[\log P(Y|Z, W)],$$

and straightforward calculations show that

$$\begin{aligned}
\mathbb{E}_Y[\log P(Z)] &= \sum_{d=1}^{D} \tau_{1d} \log q(Z_1 = d) + \sum_{t=1}^{n} \sum_{d,d'=1}^{D} \eta_{tdd'} \log \pi_{dd'}, \\
\mathbb{E}_Y[\log P(W|Z)] &= \sum_{d=1}^{D} \sum_{k=1}^{K_d} \sum_{t=1}^{n} \tau_{td} \delta_{tdk} \log \lambda_{dk}, \\
\mathbb{E}_Y[\log P(Y|Z, W)] &= \sum_{d=1}^{D} \sum_{k=1}^{K_d} \sum_{t=1}^{n} \tau_{td} \delta_{tdk} \log \phi_d(Y_t, \gamma_{dk}),
\end{aligned}$$

where $\tau_{td} = P(Z_t = d|Y)$, $\eta_{tdd'} = P(Z_t = d, Z_{t+1} = d'|Y)$, $\delta_{tdk} = P(W_t = dk|Z_t = d, Y)$.

Explicit expression of the parameter estimators is then obtained.

$$\begin{aligned}
\widehat{\pi}_{dd'} &\propto \sum_{t=1}^{n} \eta_{tdd'}, \\
\widehat{\lambda}_{dk} &= \frac{\sum_{t=1}^{n} \widehat{\tau}_{td} \widehat{\delta}_{tdk}}{\sum_{t=1}^{n} \widehat{\tau}_{td}}, \\
\widehat{\gamma}_{dk} &= \underset{\gamma_{dk}}{\operatorname{argmax}} \sum_{t=1}^{n} \widehat{\tau}_{td} \sum_{l=1}^{K_d} \widehat{\delta}_{tdl} \log \phi(y_t; \gamma_{dl}).
\end{aligned}$$

Like any EM algorithm, the algorithm behavior strongly depends on the initialization step. The naive idea of testing all the possible combinations of the components leads to intractable calculations, we chose to follow the strategy proposed by (Baudry *et al.*, 2010) based on a hierarchical algorithm. It consists in starting from a HMM with K-hidden states to finish with a HMM with D-hidden states by combining the best pair of clusters at each step. To do this, we define three likelihood-based criteria involving the Markovian dependency structure of the latent variables:

$$\begin{array}{ll} \nabla_{kl}^{Y} & = \mathbb{E}_Y \left[\log P(Y; G'_{k\cup l})\right], \\ \nabla_{kl}^{Y,Z} & = \mathbb{E}_Y \left[\log P(Y, Z; G'_{k\cup l})\right], \\ \nabla_{kl}^{Y,W} & = \mathbb{E}_Y \left[\log P(Y, W; G'_{k\cup l})\right], \end{array} \tag{2.11}$$

where $G'_{k\cup l}$ is a $G-1$ clusters[1] obtained by merging the two clusters k and l from the model with G clusters ($D < G < K$). Assuming that the hierarchical algorithm is at the G-th step, two clusters k and l are combined if they maximize one of the combining criteria ∇_{kl} defined above. Once two clusters are combined into a new cluster, a model with $G-1$ clusters is obtained. Due to the constraints applied on the transition matrix of $\{W_t\}$, the resulting estimates of the model parameters do not correspond to the maximum likelihood estimates. So, to get closer to a local maximum, a few iterations of the EM algorithm are derived to increase the likelihood of the reduced model.

Selection criteria for the number of hidden states

In many situation, D is unknown and difficult to choose. To tackle this problem, we proposed model selection criteria derived from the classical mixture framework.

The first criterion is the Bayesian Information Criterion, equal to

$$BIC(m) = \log P(Y|m, \widehat{\theta}) - \frac{\nu_m}{2}\log(n), \tag{2.12}$$

where ν_m is the number of free parameters of the model m and $P(Y|m, \widehat{\theta})$ is the maximum likelihood calculated at its maximum. However BIC is not devoted to classification and it is expected to mostly select the dimension according to the global fit of the model. Moreover BIC does not distinguish the two latent variables whereas their role differs in the emission distribution modeling. For this reason we also considered the Integrated Complete Likelihood (ICL) criterion , proposed by (Biernacki *et al.*, 2000) to select the number of clusters. Although ICL is established in the independent mixture context,(Celeux and Durand, 2008) used it in the HMM context and showed that it seems to have the same behavior.

Consider a model-based clustering with a general latent variable U, ICL criterion is based on the integrated complete likelihood and is defined by (McLachlan and Peel, 2000) as

$$\begin{array}{rcl} ICL(m) & = & \mathbb{E}_Y \left[\log P(Y, U|m, \widehat{\theta})\right] - \frac{\nu_m}{2}\log(n) \\ & = & \log P(Y|m, \widehat{\theta}) - \left(-\mathbb{E}_Y \left[\log P(U|Y, m, \widehat{\theta})\right]\right) - \frac{\nu_m}{2}\log(n). \end{array} \tag{2.13}$$

Hence ICL can be viewed as BIC with an additional penalty term, which is the conditional entropy of the hidden variable $\mathcal{H}_Y(U) = -\mathbb{E}_Y \left[\log P(U|Y, m, \widehat{\theta})\right]$. This penalty term is a measure of the uncertainty of the classification.

In our model, the latent variable U is the couple (Z, W), and a direct rewriting of (2.13) leads to

$$ICL(m) = \log P(Y|m, \widehat{\theta}) - \mathcal{H}_Y(Z, W) - \frac{\nu_m}{2}\log(n). \tag{2.14}$$

However the conditional entropy can further be decomposed as

$$\mathcal{H}_Y(Z, W) = \mathcal{H}_Y(Z) + \mathbb{E}_Y[\mathcal{H}_{Y,Z}(W)],$$

[1] the term 'cluster' refers to either a component or a mixture of components.

which gives raise to two different entropies: $\mathcal{H}_Y(Z)$ measures the uncertainty of the classification into the hidden states whereas $\mathbb{E}_Y[\mathcal{H}_{Y,Z}(W)]$ measures the classification uncertainty among the components, within each hidden state. Since this latter entropy may be irrelevant for our purpose, we derived an alternative version of ICL, where only the former entropy is used for penalization:

$$ICL_Z(m) \quad = \quad \log P(Y|m,\widehat{\theta}) - \mathcal{H}_Y(Z) - \frac{\nu_m}{2}\log(n). \tag{2.15}$$

These three criteria, BIC, ICL and, ICL_Z, display different behavior in independent mixtures and in HMM. In the independent case, the number of free parameters ν_m only depends on K and the likelihood $\log P(Y|m,\widehat{\theta})$ remains the same for a fixed K, whatever D. The criteria BIC and ICL, given in Equations (2.12)and (2.14) respectively, are thereby constant whatever the number of hidden states. Moreover, the ICL_Z always increases with the number of hidden states so none of these three criteria can be used in the independent mixture context. On the contrary, in the case of HMM, the likelihood $\log P(Y|m,\widehat{\theta})$ varies with the number of hidden states. Furthermore, because the number free parameters depends on D through the dimension of the transition matrix, the number of free parameters of a D-state HMM differs from that of a $(D-1)$-state HMM, even with a same K. For all theses reasons, these three criteria can be considered to select the number of clusters.

We performed a simulation study to evaluate the behavior of the three merging criteria (∇^Y, $\nabla^{Y,Z}$, $\nabla^{Y,W}$) and the selection criteria (BIC, ICL, ICL_Z). following the simulation design of (Baudry *et al.*, 2010). We evaluated the criteria by both the MSE (Mean Square Error) of the conditional probabilities and the classification error rate when observations are assigned into an hidden state with the *Maximum A Posteriori* rule. The criteria ∇^Y and $\nabla^{Y,Z}$ behaved similarly in most cases. Nevertheless, we proposed the ∇^Y criterion for combining the clusters due to the smaller standard deviation of its MSE and its better rate of correct classification in highly overlapping clusters. For the estimation of the number of clusters, we pointed out that ICL overestimated the number of clusters. It can be explained by the fact that ICL involves the latent variable W which is linked to the components. The criteria BIC and ICL_Z do not depend on W and are more reliable to estimate the number of clusters. According to our simulation study, we opted for ICL_Z which has the highest rates of correct estimations of the number of clusters whatever the scenario. To summarize, we propose using ∇^Y as the combining criterion and estimating the number of clusters by ICL_Z and name this strategy HMMmix summarized below:

1. Fit an HMM with K components.
2. From $G = K, K-1, ..., 1$
   - Select the clusters k and l to be combined as:

$$(k,l) = \underset{i,j\in\{1,...,K\}^2}{\operatorname{argmax}} \nabla^Y_{ij},$$

   - Update the parameters with a few steps of the EM algorithm to get closer to a local optimum.
3. Selection of the number of groups $\widehat{D}$:

$$\begin{aligned}\hat{D} &= \underset{k\in\{K,...,1\}}{\operatorname{argmax}} ICL_Z(k) \\ &= \underset{k\in\{K,...,1\}}{\operatorname{argmax}} \log P(Y,\widehat{Z}|k,\widehat{\theta}_k) - \frac{\nu_k}{2}\log(n),\end{aligned}$$

Some remarks

The interest of accounting for dependency in the hierarchical process is beneficial for the estimation of conditional probabilities. We also studied the behavior of HMMmix in the absence of dependency by generating independent observations. We observed that the performances of HMMmix were very poor. Indeed, in an HMM, the likelihood function varies when combining components, whereas it remains constant in an independent mixture. HMMmix exploits these variations both in terms of likelihood and conditional entropy and has therefore an erratic behavior when the dependency is actually absent. The performances of the method proposed by Baudry *et al.* (2010) were much better in the independent case. We therefore advise to use this approach in the independent case. In some situation, where independence is not obvious, a BIC comparison between the two methods is a solution to decide whether the dependence should be taken into account.

2.2.3 Variational Bayes approach for model aggregation

The work presented above focuses on the emission distribution in a general framework. In Volant *et al.* (2012) we considered a two-state HMM, where the emission distribution of the first state, denoted ψ_0, is completely known whereas the second emission distribution, ψ_1, is unknown and can be multimodal. Our objective was to retrieve the binary latent variable $\{Z_t\}$ associated with each observation by estimating the probability $\tau_t = P(Z_t = 0|Y)$, which requires the estimation of ψ_1. To do this, we considered a finite collection of mixtures $\mathcal{M}$ with a varying number of components and gathered information provided by each of them using Bayesian model averaging (BMA).

Bayesian model averaging, mainly developed by Hoeting *et al.* (1999), provides the general framework. Madigan *et al.* (1994), Madigan and Hutchinson (1995), Raftery *et al.* (1997); Volinsky *et al.* (1997) or Raftery *et al.* (2003) demonstrated that BMA can improve predictive performances and parameter estimation. Jaakkola and Jordan (1998) also demonstrated that model averaging often provides a gain in terms of classification and fitting.

Let $\mathcal{M}$ a collection of models and $\Delta = \delta(\Theta)$ a contrast for which we aim at estimating is conditional distribution:

$$\begin{aligned} P(\Delta|Y) &= \sum_{m\in\mathcal{M}} P(\Delta, M = m|Y) \\ &= \sum_{m\in\mathcal{M}} P(\Delta|M = m, Y)P(M = m|Y) \\ &= \sum_{m\in\mathcal{M}} \alpha_m P(\Delta|M = m, Y) \end{aligned}$$

Consequently when averaging, each model of the collection is weighted and the determination of these weights α_m is the key ingredient of all these approaches. The calculation of α_m requires to integrate the joint conditional distribution $P(M, \Theta|Y)$, where Θ is the random vector of model parameters. Several approaches can be used: a Laplace approximation of this integral leads to the BIC criterion, which is questionable for small sample sizes. One other commonly used method is MCMC (Monte Carlo Markov Chain, Andrieu (2003)) which samples the distribution and provides an accurate estimation of the joint conditional distribution, but at the cost of huge computational time.

In the unsupervised classification context, the problem is even more difficult as we need to integrate the conditional $P(M, \Theta, Z|Y)$ since the latent variables $\{Z_t\}$ are unobserved. This distribution is generally not tractable but, for a given model, Beal and Ghahramani (2003) developed a variational Bayes strategy to approximate $P(\Theta, Z|Y)$ in the exponential family framework. We used this strategy to propose variational-based weights for model averaging in a binary HMM model. We proved that the proposed weights are optimal in terms of Kullback-Leibler divergence from the true conditional distribution $P(M|Y)$. To this end, we optimized

the KL-divergence between $P(\Theta, Z, M|Y)$ and an approximated distribution $Q_{\Theta,Z,M}$. We also derived two other weights based on the approximated distribution of $P(\Theta, Z|M, Y)$. Finally we proposed a complete inference procedure for the case of binary classification when a collection $\mathcal{M}$ of mixtures of parametric exponential family distributions was considered to estimate ψ_1. We used this framework to discuss the accuracy of the weights and the efficiency of the averaged estimators of $\{\tau_t\}$.

Variational weights

The aim of model averaging is to account for the information of each model of the considered collection. It implies evaluating the conditional distribution:

$$P(M|Y) = \int P(H, M|Y)dH, \tag{2.16}$$

where $H = (Z, \Theta)$ stands for all hidden variables and M denotes a model. To calculate this distribution, we need to compute the joint posterior distribution of H and M. Due to the latent structure, this is not feasible. However, the variational theory allows one to derive an approximation of this distribution. It is based on the minimization of the KL-divergence between $P(H, M|Y)$ and an approximated distribution $Q_{H,M}$ and simple algebraic calculations show that $KL(Q_{H,M}||P(H, M|Y))$ minimization with respect to $Q_{H,M}$ can be separated in two parts, which can be realized independently:

$$\min_{Q_M} \left[KL(Q_M||P(M|Y)) + \sum_m Q_M(m) \min_{Q_{H|m}} KL(Q_{H|m}||P(H|Y, m)) \right]. \tag{2.17}$$

Since we are mostly interested in Q_M which provides an approximation of $P(M|Y)$, we first minimized the KL-divergence with regard to Q_M leading to weights that depend on $Q_{H|m}$. We showed that the optimal weights, minimizing $KL(Q_{H,M}||P(H, M|Y))$ with respect to Q_M, for given distributions $\{Q_{H|m}, m \in \mathcal{M}\}$ are proportional to

$$\overline{\alpha}_m(Q_{H|m}) \propto P(m) \exp[-KL(Q_{H|m}||P(H|Y, m)) + \log P(Y|m)], \tag{2.18}$$

with $\sum_m \overline{\alpha}_m(Q_{H|m}) = 1$. If $Q_{H|m} = P(H|Y, m)$ then $\overline{\alpha}_m$ resumes to $P(m|Y)$.

We derived three different weights based on a variational Bayes approximation. The first one comes from the complete optimization of the KL divergence. The second one is based on a plug-in approach. The third weight uses the variational posterior as a proposal for importance sampling. The following weights are general and can be used to estimate any parameter common to all the models of the collection.

Full variational approximation To solve the optimization problem (2.17), for each model m of the collection $\mathcal{M}$, we minimized the divergence $KL(Q_{H|m}||P(H|Y, m))$ with respect to $Q_{H|m}$ by using the Variational Bayes EM algorithm (VBEM; Beal and Ghahramani (2003)). If $P(Y, Z|\Theta, M)$ belongs to an exponential family and $P(\Theta|M)$ is its conjugate prior, then this algorithm minimizes $KL(Q_{H|m}||P(H|Y, m))$ within the class of factorized distributions $\mathcal{Q}_m = \{Q_{H|m} : Q_{H|m} = Q_{Z|m}Q_{\Theta|m}\}$ and provides an approximate of $P(H|Y, m)$ defined as:

$$Q^{VB}_{H|m} = \arg \min_{Q \in \mathcal{Q}_m} KL(Q_{H|m}||P(H|X, m)).$$

We used it to define variational weights $\widehat{\alpha}^{VB}_m$ achieving the optimization problem (2.17) for factorized conditional distribution $Q_{H|m}$. These variational weights are derived from Equation (2.18) and are defined as:

$$\widehat{\alpha}^{VB}_m \propto P(m) \exp[- \min_{Q_{H|m} \in \mathcal{Q}_m} KL(Q_{H|m}||P(H|X, m)) + \log P(X|m)].$$

Plug-in weights The weights $\alpha_m = P(M = m|Y)$ can be estimated by using a plug-in estimation based on a direct application of Bayes theorem. By definition $P(m|Y) = P(m)P(Y|m)$ and is equal to $P(Y|m,\Theta)P(\Theta|m)/P(\Theta|Y,m)$ for any value of Θ. Using the distribution $Q^{VB}_{\Theta|m}$ resulting from the VBEM algorithm as an approximation of $P(\Theta|Y,m)$ and setting Θ at its (approximate) posterior mean $\theta^* = \mathbb{E}_{Q^{VB}_{\Theta|m}}(\Theta)$, the plug-in weight of model m was defined as:

$$\widehat{\alpha}^{PE}_m \propto P(m)\frac{P(Y|m,\theta^*)P(\theta^*|m)}{Q^{VB}_{\Theta|m}(\theta^*)}. \tag{2.19}$$

Importance sampling The weights α_m can be also estimated via importance sampling (Marin and Robert (2009)) such that

$$P(m|Y) \propto \int P(m)\frac{P(Y|h,m)P(h|m)}{R(h)}R(h)dh$$

for any distribution R. Importance sampling provides an unbiased estimator of $P(m|Y)$ and the choice of the importance function R impacts the variance of the estimator. It is known that the minimal variance is reached when $R(H)$ equals $P(H|Y)$.

In the variational framework, the approximated posterior distribution $Q^{VB}_{H|m}$ is a natural choice for the importance function R, leading to the following weights:

$$\widehat{\alpha}^{IS}_m \propto P(m)\frac{1}{B}\sum_{b=1}^{B}\frac{P(Y|H^{(b)},m)P(H^{(b)})}{Q^{VB}_{H|m}(H^{(b)})}, \qquad \{H^{(b)}\}_{b=1,\dots,B} \text{ i.i.d. } \sim Q^{VB}_{H|m}.$$

Although this estimate is unbiased, when the number of observations is large, it may require a long computational time to get a reasonably small variance.

Application in a binary hidden Markov model

A two-state hidden Markov model (HMM) is considered. The $\{Y_t\}$ are supposed to be independent conditionally to the latent variables $\{Z_t\} \in \{0,1\}$, distributed according to a first order Markov chain. The function ψ_0 is a known parametric function whereas ψ_1 is unknown. To approximate this latter, we considered a finite collection of mixture $\mathcal{M} = \{f_m; m = 1,\dots,M\}$, where

$$f_m(.) = \sum_{k=1}^{m}\eta_k\phi(.;\gamma_k),$$

with η_k is the mixing proportion of the k-th component from cluster d ($\forall k \in \{1,...,m\}$, $0 < \eta_k < 1$ and $\sum_k \eta_k = 1$) and $\phi(.;\gamma)$ denotes a parametric distribution belonging to the exponential family known up to the parameter vector γ. Since ψ_1 is approximated by a mixture, the initial binary HMM with latent variable Z can be rephrased as an $(m+1)$-state HMM with hidden Markov chain $\{W_t\}$ taking its values in $\{0,\dots,m\}$. Its transition matrix deduced from (2.10) equals:

$$\Omega_m = \begin{pmatrix} \pi_{00} & \pi_{01}\eta_1 & \dots & \pi_{01}\eta_m \\ \pi_{10} & \pi_{11}\eta_1 & \dots & \pi_{11}\eta_m \\ \vdots & \vdots & \vdots & \vdots \\ \pi_{10} & \pi_{11}\eta_1 & \dots & \pi_{11}\eta_m \end{pmatrix},$$

where $(\pi_{00},\pi_{01},\pi_{10},\pi_{11})$ are the terms of the transition matrix of $\{Z_t\}$. In this framework, we proposed a variational-based approach to provide an aggregated estimator of $\{\tau_t\}$, where $\tau_t = P(Z_t = 0|Y)$. First for each $m \in \mathcal{M}$, VBEM algorithm is applied to provide an approximation of $P(H|Y,m)$, i.e. $Q^{VB}_{H|m}$. Inference is made on the $(m+1)$-state hidden Markov model involving W rather than the binary hidden Markov model involving Z, because $P(Y,Z|\Theta,M)$ does not belong

to the exponential family whereas $P(Y, W|\Theta, M)$ does. Second, weights, $\widehat{\alpha}_m^{VB}$, $\widehat{\alpha}_m^{PE}$ and, $\widehat{\alpha}_m^{IS}$ are computed. Based on these weights, an averaged estimate of the distribution ψ_1 is derived:

$$\widetilde{\psi_1}^{\mathcal{A}} = \sum_{m \in \mathcal{M}} \widehat{\alpha}_m^{\mathcal{A}} \widehat{f}_m,$$

where $\mathcal{A}$ corresponds to one of the proposed approaches (VB, PE or IS). The estimation of the posterior probability τ_t is similarly defined by its averaged estimate:

$$\widetilde{\tau}_t^{\mathcal{A}} = 1 - \sum_m \widehat{\alpha}_m^{\mathcal{A}} \mathbb{E}_{Q_{W|m}^{VB}}(Z_t),$$

where $\mathbb{E}_{Q_{W|m}^{VB}}(Z_t)$ corresponds to the expected value of Z calculated with the optimal variational posterior distribution of W.

In a simulation study, we studied the accuracy of $\widehat{\alpha}^{VB}$ and $\widehat{\alpha}^{PE}$ in terms of weight estimation. The variational and plug-in weights were compared to the importance sampling, considered as a reference as it provides an unbiased estimate of the true weights whatever the choice of R. It was done by calculating the total variation distance. Similarities between the importance sampling method and the variational method clearly appeared. The variational method provides a good estimation of the true weights in contrast to the plug-in weights. We studied also the accuracy of the posterior probability estimator by studying the Mean Squared Error (MSE) of aggregated estimators based on the weights. The variational weights provide better results in terms of MSE and its standard deviation than do the plug-in weights. These results are very close to those of the importance sampling weights and even often better. The misclassification rate confirms the closeness between our approach and the importance sampling. Both methods have a quite similar misclassification rates whatever the simulation condition. Hence, when the computational time of the importance sampling method becomes very high, we get a real advantage by using the variational weights.

We analyzed epidemiologic and transcriptomic datasets. The aggregation model refines the estimation of posterior probabilities, in particular the classification is different in cases where the probability is close to 0.5, i.e. when the classification is difficult.

2.2.4 Conclusion and perspectives

In our works, HMM are viewed as a generalization of the classical mixture model where the neighborhood structure was accounted for via a Markov dependency between the latent variables. It is a useful model for analyzing genomic data generated from array where the probes covered the whole genome. Usually the emission distributions are modeled with a parametric distribution leading to tractable calculations but to a poor fit, when the distribution of the data is multimodal. To address this problem, we proposed a general modeling of an HMM where the emission distributions are mixture of parametric distributions to get more flexibility. The two works are examples of what we can do in such models and reveal a diversity of computational and methodological problems.

In the work on the estimation of the emission distributions in a general HMM, we proposed a hierarchical combination to initialize the estimation algorithm. We proposed three different criteria for combining the clusters and three different criteria for model selection. Based on a simulation study, we selected ∇^Y as combining criterion and ICL_Z for the estimation of the number of clusters. This strategy provides estimation of the conditional probabilities close to the true ones and they are very robust in terms of MSE whatever the dependency. Although the method is described with an unknown number of hidden states D and is illustrated with mixture of Gaussian distributions, the same approach can be used when D is known or with other parametric distribution family in the mixtures. For the initial number of components K, a brief simulation study has shown that for a large enough value of K, the classification still remains the same. Further work would be needed to derive a computationally efficient and theoretically grounded method to choose K.

A remaining problem of HMMmix is the computational time, especially when the size of the dataset is greater than 10000. This is due to the calculation of the ∇^Y criterion which is linked to the observed log-likelihood. The computation of this observed log-likelihood requires the forward loop of the forward-backward algorithm whose complexity is linear in the number of observations. Otherwise, the number of models considered in the hierarchical procedure is $\sum_{d=D}^{K} d(d-1)/2$, so the computational time dramatically increases with K and is of order $O(nK^3)$. Consequently, to decrease the computational time, the space of models to explore should be reduced.

In the binary hidden Markov model, we proposed averaged estimators of the posterior probability within a variational Bayesian framework. This approach allows us to avoid model selection and take model uncertainty into account. Moreover it does not required more computational time than the more commonly used selection approach. The results obtained on simulated data show that our method enhances the estimator in terms of MSE in many simulation conditions. The averaging approach improves the posterior probability estimation compared to the classical selection approach. Moreover, we show that optimal variational weights are close to importance sampling weights. Since the importance sampling copes with computational time problems for high dimensional datasets, our method is of significant interest in this case. On real data, the aggregation model still refines the estimation of posterior probabilities, especially in the neighborhood of 0.5, where it is difficult to take a decision about the classification.

This method of averaging is general and can be applied to estimate any common parameter available in each model of the collection while mixtures of parametric exponential family distributions are considered and conjugate priors exist. The Bayesian framework requires to specify hyperparameters that could impact the results. Stevenn Volant studied their impact (see Chapter 4.5 of his PhD). He showed that the main contribution of the hyperparameters is in the normalization constant of the emission distribution. If the mixture used to approximate the emission distributions has components with homogeneous variance, hyperparameters have less impact on the weight calculation than hyperparameters of a mixture with components with heterogeneous variance. He illustrated this fact in a simulation study and concluded that an homogeneous mixture should preferred.

To make our models available to the broader community, the two models are available in R packages. For an estimation of the emission distributions with mixture Caroline Bérard and Stevenn Volant made available an R package HMMmix on the CRAN. This package proposes to reduce a K-state HMM to a D-state HMM. For the binary hidden Markov model with a fixed Gaussian distribution as emission for one state and a mixture of Gaussian for the other state, Stevenn Volant made available an R package named VBMA4hmm (Variational Bayes Models Averaging for hidden Markov models) available on the CRAN. Optimal variational weights and averaged estimator of the posterior probabilities are calculated.

2.3 Classification

The works presented in this section correspond to results of the internship and the PhD of Caroline Bérard

2.3.1 Background

Model-based clustering aims at classifying observations in clusters. Once the model is selected and the parameters are estimated, the probability of $\tau_{ki} := P(Z_i = k|Y)$ for each observation is available. It is called the conditional probability in a frequentist framework or the posterior probability in a Bayesian framework. It is used to assign observation into clusters.

The classification is usually performed at the observation level whereas the observations are not always relevant for the interpretation. A genomic example is the tiling array data: observations are the probes and the biologist is mainly interested by the classification of a subset of probes describing a biological entity like a gene. In this case, when region defined by adjacent

observations, the most commonly used method is a sliding window approach where observations are merged *a priori*. In Bérard *et al.* (2011), we proposed a conditional probability for a given subset of observations with arbitrary structure and a procedure of region classification.

By definition, a classification rule is a function that maps $\mathcal{Y}$ into $\{1, ..., K\}$, where K is the number of modality of the latent variable Z. The most popular one is the *Maximum A Posteriori* (MAP) classification rule, defined as

$$\underset{k=1,...,K}{\operatorname{argmax}} \ \tau_{ki}$$

It is known to minimize the prediction error. Nevertheless it does not prevent against a high level of misclassification and it does not take into account that only a small number among the K populations may be of major importance for the experimenter. To circumvent these drawbacks, a strategy is to focus on the classification of observations in regions of $\mathcal{Y}$ where populations do not strongly overlap. It is illustrated by the thresholded classification rule

$$\begin{cases} p & \text{if } \tau_{pi} > 1 - \alpha, \\ 0 & \text{(not classified) otherwise,} \end{cases}$$

where $0 < \alpha < 1/2$ is a parameter to be chosen by the experimenter. Alternatively, one can take into account the asymmetry between populations by selecting unequal misclassification cost functions that emphasize the cost of misclassification $c_{k\ell}$ from a non-interesting population ℓ into a population k of interest. This has been investigated in McLachlan and Peel (2000) for instance. Although the use of these alternative rules is attractive, there exist very few guidelines about how to tune parameters α or $c_{k\ell}$ and what are the consequences of this tuning on the misclassification rate. In Martin-Magniette *et al.* (2008a) we proposed a statistical procedure to control the proportion of observations wrongly classified in one of the two populations of a mixture. These two contributions are described in the following.

2.3.2 Conditional probabilities for a subset of observations in a HMM

In a classical HMM, the conditional probabilities for an observation are obtained as a by-product of the Forward-Backward algorithm. Nevertheless, an observation may not be the relevant biological entity. In Bérard *et al.* (2011) we proposed a generalization of the conditional probabilities for a region, defined as a set of observations that can be decomposed into sub-regions of adjacent observations. The conditional probability for such a region g to belong to group k is defined as the probability for all its probes to belong to group k:

$$Q_{gk,Y} := P(\forall t \in g, Z_t = k|Y).$$

The example in mind being the gene structure, to explain the calculation, the sub-regions are named 'exons' and to the spaces between them as 'introns'. Consider a gene g with Q exons (and $Q-1$ introns), denote e_q the position of the first probe of exon q and i_q the position of the first probe of intron q; thus $i_q - 1$ refers to the last probe of exon $q-1$. As convention, i_Q denotes the position of the first probe after the end of the gene. If Y_u^v denotes $\{Y_t\}_{u \le t \le v}$, the conditional probability $Q_{gk,Y}$ equals:

$$\begin{aligned} P(Z_{e_1} = k|Y_1^{e_1}) \quad \times \quad & \left(\prod_{t=e_1+1}^{i_1-1} A_{k,t} \right) \times \prod_{q=2}^{Q-1} \left(B_{k,q} \times \prod_{t=e_q+1}^{i_q-1} A_{k,t} \right) \\ \times \quad & B_{k,Q} \times \left(\prod_{t=e_Q+1}^{i_Q-2} A_{k,t} \right) \times P(Z_{i_Q-1} = k|Z_{i_Q-2} = k, Y_{i_Q-1}^n) \end{aligned}$$

where $A_{k,s} = P(Z_s = k|Z_{s-1} = k, Y_s)$ and $B_{k,q} = P(Z_{e_q} = k|Z_{i_{q-1}-1} = k, Y_{i_{q-1}}^{e_q})$, calculated with the Forward recursion of the Forward / Backward algorithm. Figure 2.3.2 illustrates a region and the terms $A_{k,s}$ and $B_{k,q}$.

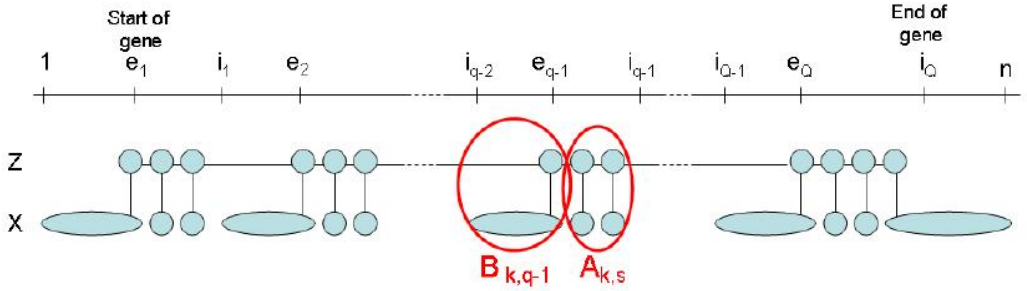

Figure 2.6: Schematic representation of a region

These conditional probabilities are not directly usable because their sum over $k \in \{1, ...K\}$ is not equal to one. We thus proposed a classification procedure. First region homogeneity is verified using a *unistatus* value. It is based on the log-ratio $\log(\sum_k Q_{gk,Y} / \sum_k Q_{gk})$, where $Q_{gk} = P(\forall t \in g, Z_t = k)$ is the non-conditional version of $Q_{gk,Y}$. Because $Q_{gk,Y}$ goes to zero for long regions and the ratio with Q_{gk} does not correct this effect, we applied an additional linear correction on the $\log(\sum_k Q_{gk,Y} / \sum_k Q_{gk})$ with respect to the length of the sub-regions and their number to define a corrected *unistatus* value which is independent of the length of the region and its composition. Then if this corrected *unistatus* is greater than 0, the region is considered homogeneous and classified according to the MAP rule by using the conditional probabilities $Q_{gk,Y} / \sum_{l=1}^{K} Q_{gl,Y}$.

We applied this region classification on tiling array data but it was difficult to interpret biologically the results given the fact that the functional annotation is still unclear.

2.3.3 Control of the false-positives in a mixture of two linear regressions

In Martin-Magniette *et al.* (2008a), we devised a mixture of two linear regressions to investigate protein-DNA interactions. Observations are the hybridization signal of probes of a microarray and the mixture model aims at distinguish probes covering DNA regions where the protein of interest is absent and probes covering DNA regions where the protein of interest is crosslinked to the DNA. This model is explained in Section 3.3.2. If we used the MAP rule, it implicitly means that misclassifications in both groups have the same cost. But from a biological point of view, the cost is not the same since only the second group is relevant and any observations wrongly assigned in this group mislead the biologist who looses time on a wrong interpretation. In conclusion, even in model-based clustering, when only some groups are of interest, false positives are of concern.

In the hypothesis test theory, the false discovery control is performed by controlling the probability to reject wrongly the null hypothesis. We proposed an analogous concept in the mixture model framework by controlling the probability for an observation to be wrongly assigned to the group of interest. Consider n independent random variables $\{Y_i\}$ and assume that the distribution of Y_i conditional to a given covariate x_i is a mixture of two linear regressions. Let $Z_i \in \{0, 1\}$ is a binary latent variable indicating the observation status. The conditional distribution of Y_i for a given x_i is

$$\pi\phi_0(Y_i|x_i) + (1 - \pi)\phi_1(Y_i|x_i),$$

where π is the mixing proportion and $\phi_j(\cdot|x)$ stands for the probability density function of a Gaussian distribution with mean $a_j + b_j x$ and variance σ^2.

We focused on the classification task with objective to control the wrongly assign observations in one component. Let $\tau_i := P(Z_i = 1|x_i, Y_i)$, we proposed to control the probability to be

wrongly assigned to the group of interest by fixing a threshold s such that

$$P\{\tau_i > s \mid x_i, Z_i = 0\} = \alpha,$$

where α is a predefined level. By definition, the threshold s depends on on α and x_i and has an explicit form in a mixture of two homoscedastic linear regressions. Moreover straightforward calculations show that $P\{\tau_i > s \mid x_i, Z_i = 0\}$ is equivalent to $P\{Y > Y_i \mid x_i, Z_i = 0\}$, where Y is distributed according to a Gaussian of mean $a_0 + b_0 x_i$ and variance σ^2. This latter probability being uniformly distributed on $[0, 1]$, it has allowed one to control a misclassication rate analogous to the false discovery rate of the hypothesis theory Benjamini and Hochberg (1995).

2.3.4 Conclusion and perspectives

The use of model-based clustering in biological projects obliges to think about the classification. When the observations are less relevant than a given subset of observations, it could be interesting to classify directly the subset. We showed that in this case, it is necessary to generalize the definition of the conditional probability and to be vigilant when subset are classified since the sum the new conditional probabilities is not equal to one. From a practical point of view, the subsets should be given *a priori* from external knowledge and this latter point is sometimes questionable for the biologists.

When there exists an asymmetry in the group interpretation, we are confronted to false-positives. We showed that in binary mixtures, it is sometimes possible to control the probability to be wrongly assigned and to control the misclassification rate. In binary HMM, Sun and Cai (2009) proposed a FDR control. It is based on a monotone condition on the ratio of the emission distributions. When the number of groups is greater than 2, the problem becomes more difficult because the notions of false-positives should be redefined. In an ongoing project in collaboration with Tristan Mary-Huard (UMR AgroParisTech/INRA MIA 518), we are extending the idea of false-positives when P groups are of interest among K (Mary-Huard *et al.*, 2013). We build classification rules that guaranty that as many observations as possible are classified in the groups of interest, under the constraint that the proportion of misclassified observations in these groups is controlled. As such, our idea can be understood as an extension of the BFDR developed in Efron and Tibshirani (2002) to general finite mixture models.

Chapter 3

Statistical approaches dedicated to genomic questions

3.1 Context

The works presented in this chapter are dedicated to the data analysis of microarray. Figure 3.1 describes the principle based on hybridizations. Microarrays are miniaturized biological devices consisting in DNA molecules named the probes, that are arranged at a microscopic scale onto a solid support. In a second step, mRNA are extracted from the biological samples under study, stabilized through a reverse transcription into cDNA and labeled with fluorescence dyes. They are then deposited on the support. The labeled samples as well as the probes being single DNA strand, as soon as a target sequence is the complementary of a probe sequence, both elements hybridize. The next step consists in washing, drying the slide to keep only the targets fixed to probes. Finally a scanner records, after excitation of the two fluorochromes at given wavelengths, the intensity of the fluorescence emission signals, that is proportional to transcript levels in the biological samples.

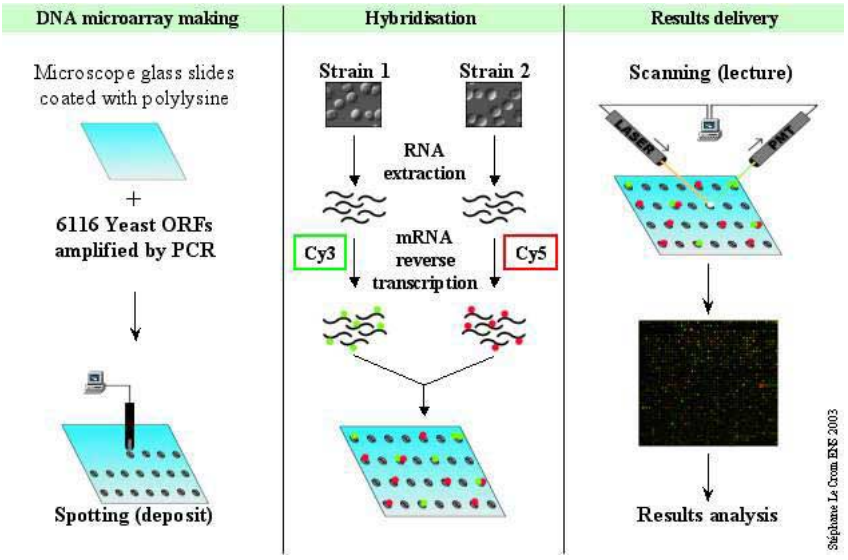

Figure 3.1: Principle of a microarray experiment

Microarray technology revolutionized the molecular biology and various questions not always well-defined appeared. Statistical approaches dedicated to microarray are mainly based on a good general culture in statistics (linear model, mixture, multiple tests, ...) and pragmatism. The first section presents modeling questions to remove the technical biases that corrupt the

hybridization signals. The second section is devoted to model-based clustering methods developed for specific questions. These methods were applied in various genomic projects and the biological results are described in the next chapter.

3.2 Signal modeling for microarray data normalization

The works presented in this section are the results of a collaboration with Julie Aubert and Jean-Jacques Daudin for the statistical part and with the persons of the transcriptomic platform of URGV and Eric Cabannes (Institut Pasteur) for the biological aspects.

3.2.1 Background

Biases exist at each step of the protocol from the sample preparation to the gene expression measurement. Consequently, normalization is a necessary step and consists of identifying and correcting technical biases on the observed intensities of the probes to deduce the value of the biological signal. Normalization is impossible to circumvent and has an impact on the sequel of the statistical analysis.

Even if the statisticians are not the most competent person to track these biases, in practice they often do this task, since they have the possibility to propose a systematic approach based on the modeling to define, quantify and correct the biases. Working on normalization is like an investigation of a private detective: We observe the technical biases on the gene expression measurements, have to inform the biologists that the quality of their data is bad and find the origin of the biases or devise a procedure to remove them.

Between 2003 and 2005, we were several members of the team "Statistique and Genome" of UMR AgroParisTech/INRA MIA 518 working on biases occurring in microarray experiments. It allowed us to have a good knowledge of these biases and to advice biologists. Working on technical biases requires to discuss a lot with the biologists generating the data to really understand each step of the protocol. It is also important to have the confidence of the biologists. Indeed applying a rigorous protocol takes time and if the biologists are not aware of the importance of some steps, they can modify them to speed up the generation of the data but also the generation of the technical biases. Moreover technical biases are often visible on data of bad quality that biologists usually prefer to put in the trash instead of sharing them.

Briefly speaking, technical biases fall into two categories: they are either controllable and can be avoided by good experimental practice or, uncontrollable and inherent to the technology. For the latter, statistical methods are required and it begins by modeling the hybridization signals by an analysis of variance to quantify the most important biases by comparing the mean squares of each effect. If Y_{ijkg} denotes the logarithm base 2 of the measurement for array i, dye j, RNA sample k and gene g, the signal is usually modeled as follows:

$$Y_{ijkg} = \mu + A_i + D_j + V_k + G_g + (AG)_{ig} + (VG)_{kg} + (DG)_{jg} + E_{ijkg}$$

where A_i is the i^{th} array effect, D_j is the j^{th} dye effect, V_k is the k^{th} RNA sample effect, G_g is the g^{th} gene effect, and $(AG)_{ig}$, $(DG)_{jg}$ and $(VG)_{kg}$ are the corresponding interaction terms. The terms E_{ijkg} represent independent random errors with mean 0.

Kerr *et al.* (2002) showed that the most important bias is the dye bias. It comes from a incorporation difference of the two dyes $Cy5$ and $Cy3$ used to label the samples and from a natural green fluorescence of the probes. To correct the dye bias, the most popular method is the one proposed by Yang *et al.* (2002). It consists of modeling the expression differences by a function of the averaged intensities and of estimating this function by the lowess procedure (Cleveland, 1979).

In Martin-Magniette *et al.* (2005b), we were interested in the gene-specific dye bias. We wrote an analysis of the variance taken this bias into account and showed that it can be quantified thanks to self-self hybridization slides. We showed that the gene-specific dye bias is the

second major source of experimental variability and that it may alter the conclusions about the differentially expressed genes. This work is described in Section 3.3.

Most transcriptomic project are made using one or two samples on the same microarray. But improvements in dye-labeling and in the scanner sensitivity allow hybridization up to four samples simultaneously. Before the arrival of high throughout sequencing technology, the triple-target or four-target technology was considered as very promising, since it allows more flexibility in the design of experiments and a scaled down number of microarrays compared to two-color microarray. Nevertheless adding a third or fourth labeled samples on a microarray introduces new technical biases and it is important to think about the normalization of such microarray. In 2005, in the framework of an ANR project, named AgriArray, I worked with the transcriptomic platform of URGV and colleagues of UMR AgroParisTech/INRA MIA 518 to identify new biases related to the additional dyes. We proposed a normalization strategy for the multiple-target micro arrays. Thanks to an adapted modeling of the signal, we evaluated the dye bleeding and proposed a method to correct it. We also generalized the lowess procedure to correct the global dye bias. This work is described in Section 3.2.4, was published in Martin-Magniette *et al.* (2008b) but was never applied since the multiple-target technology was abandoned for the high throughput sequencing technology. It is a risk that sometimes occurs in genomics because the technologies are improved regularly.

3.2.2 Definition of an index to evaluate the gene-specific dye bias

In a self-self experiment, the same sample is hybridized twice on the same microarray. The goal of such experiment is to evaluate the impact of the technical biases on the differential analysis. During an analysis of two self-self hybridization slides, we were surprised to obtain many differentially expressed genes. The reason was that for calculating the test statistic, the quantity $(Y_{11g}+Y_{21g})-(Y_{12g}+Y_{22g})$ was wrongly calculated in place of $(Y_{11g}+Y_{22g})-(Y_{12g}+Y_{21g})$, where Y_{ijg} denotes the logarithm base 2 of the measurement for array i, dye j and, gene g[1]. With the mean $(Y_{11g}+Y_{22g})-(Y_{12g}+Y_{21g})$, no differentially expressed genes were obtained, as it is expected. We were amazed by the importance of the effect of a simple reverse of dye and decided to write the signal model to understand the phenomenon.

Model allowing for gene-specific dye bias An experiment with p dye-swaps[2] is considered and the hybridization signal is decomposed according to an analysis of variance. Let Y_{ijkg} be the logarithm base 2 of the measurement for array i, dye j, RNA sample k and gene g. We considered the following model:

$$Y_{ijkg} = \mu + A_i + D_j + V_k + G_g + (VG)_{kg} + (DG)_{jg} + E_{ijkg}$$

where A_i is the ith array effect, D_j is the jth dye effect, V_k is the kth RNA sample effect, G_g is the gth gene effect, and $(DG)_{jg}$ and $(VG)_{kg}$ are the corresponding interaction terms. The terms E_{ijkg} represent independent random errors with mean 0. If the RNA sample $k = 1$ is labeled with the dye $j = 1$ in the first array, then the observed difference of expression between the two RNA samples on the array i equals

$$Z_{ig} = V_1 - V_2 + (-1)^i(D_1 - D_2) + (VG)_{1g} - (VG)_{2g} + (-1)^i\{(DG)_{1g} - (DG)_{2g}\} + \tilde{E}_{ig},$$

where the errors $\tilde{E}_{ig}$ are independent random variates with mean 0. According to Kerr *et al.* (2002), the lowess procedure (Yang *et al.* (2002)) suppresses the first four constant terms and alleviated the DG terms and, does not alter the VG terms. Consequently, the normalized difference of expression between the two RNA samples on the array i can be decomposed as the following:

$$Z'_{ig} = \delta_g + (-1)^i\beta_g + F_{ig}, \tag{3.1}$$

[1] k index k is omitted since condition is a factor of only one level in a self-self experiment

[2] a dye-swap experiment consists of two technical replicates where opposite dye orientations are used. Thus each sample is labeled with each dye.

where $\delta_g = (VG)_{1g} - (VG)_{2g}$ is the true difference of expression between the two RNA samples and $\beta_g = (DG)'_{1g} - (DG)'_{2g}$ represents the gene-specific dye bias. When it is non null, it states that the probe corresponding to the gene g incorporates one of the dyes preferentially. The errors F_{ig} are random variates with mean 0 and are assumed to be independent despite the weak structural dependence (of order $1/G$) that exists by construction. By considering two self-self experiments, it guarantees that the true difference of expression is null and thus the model (3.1) becomes a one-way anova model:

$$Z'_{ig} = (-1)^i \beta_g + F_{ig}, \text{ for } i = 1, 2. \tag{3.2}$$

Thanks to this signal decomposition, we pointed out that it is easy to evaluate the gene-specific dye bias and to perform an analysis of this bias by testing the null hypothesis $\{\beta_1 = \ldots = \beta_G = 0\}$ against the alternative hypothesis {At least one gene is such that $\beta_g \neq 0$}. Since the number of genes, G, being large, the power of the test is high and the null hypothesis may often be rejected. For that reason, we proposed the associated test statistic as a global index to evaluate the gene-specific dye bias. We named it the Label Bias Index (LBI) and it is defined by:

$$LBI = \frac{\sum_{g=1}^{G} (Z'_{1g} - Z'_{2g})^2}{\sum_{g=1}^{G} (Z'_{1g} + Z'_{2g})^2}. \tag{3.3}$$

We proposed to compare the LBI to the expectation of the Fisher distribution, equal to $(G-1)/(G-3) \sim 1$. If the LBI is greater than 1, we conclude that the gene-specific dye bias seems important.

Application of the LBI We calculated the LBI from 11 self-self experiments from human and *Arabidopsis thaliana* samples (Table 3.1). LBI is always greater than 1 but the gene-specific dye bias is more important in the human experiments than in the *Arabidopsis thaliana* ones. For human experiment, two types of microarray are available and for one microarray type, the LBI is really high whereas the self-self experiment are made from the same sample. We also calculated the correlations between all the $\hat{\beta}_g$ for each human/array 1 experiment. They are comprised between 0.45 and 0.81. These results suggests that the dye bias may be attributed the nucleic composition of the probe and the spotting effect (see Mary-Huard *et al.*, 2004). Besides from four slides where the same sample of *Arabidopsis thaliana* had been hybridized against itself, we calculated the LBI on the six possible pairs of slides. The associated LBI was between 1.12 and 1.26, which proves its robustness. We did not investigate the origin of the gene specific dye bias but advocated in our paper to perform dye-swaps if the gene specific-dye was high.

We tested the nullity of the coefficients β_g in the model (3.2) by using Varmixt, which identifies clusters of probes with equal variance (Delmar *et al.*, 2005). Many genes are detected to have a gene-specific dye bias in the human experiments (column (b) of Table 3.1) and these results are in concordance with the LBI values. Furthermore, all the genes having a significant specific dye bias are classified in the highest variance group by Varmixt. This suggests that many genes from the highest variance group can not be detected as differentially expressed only because their variability is increased by a specific dye bias effect. It confirms that the presence of gene-specific dye bias can increase the false negative rate and so decrease the power of detection. Indeed if the gene-specific dye bias is neglected in the model (3.1) by setting β_g equal to 0 for $g = 1, \ldots, G$, then the variance of the expression difference is overestimated.

Table 3.1 contains also the mean, minimal and maximal values of the $\hat{\beta}_g$ for the detected genes. We can see that the gene-specific dye bias may multiply or divide the ratio by a factor greater than 2, which is sizeable. An analysis on the intensity level of the genes with a high specific dye bias shows that the intensity of these genes is in a large range between 5.5 and 15.7, with a median value between 9.5 and 10.2. Consequently, all expressed genes can be affected by a specific dye bias whatever their intensity level. Moreover it means that to reveal its specific dye bias, a gene needs to be transcribed.

Organism / array	Sample	LBI	(a)	(b)	Mean LR	Min. LR	Max. LR
human/array 1	sample1 t1	4.64	0	120	0.87	-1.19	1.58
human/array 1	sample2 t2	5.68	0	153	0.45	-1.46	1.52
human/array 1	sample3 t3	4.07	0	2	1.97	1.73	2.21
human/array 1	sample1 t4	4.86	0	113	1.19	-1.16	1.81
human/array 1	sample2 t5	4.64	0	33	0.19	-1.61	1.36
human/array 1	sample3 t6	6.42	0	189	1.42	-1.26	2.95
human/array 2	sample3 t2	10.29	0	8	-1.11	-1.35	-0.98
human/array 1	sample3 t2	5.15	0	3	-2.05	-2.85	-1.45
At/CATMA	leaf	1.79	0	0	-	-	-
At/CATMA	bud	1.17	0	0	-	-	-
At/CATMA	bud	1.24	0	0	-	-	-

Table 3.1: t=time, LBI = Label Bias Index, At = *Arabidopsis thaliana* (a): Number of genes differentially expressed, (b): Number of genes having a significant dye bias ($\beta_g \neq 0$), LR = Log Ratio for genes having a significant dye bias.

Comments on the LBI Few weeks after the publication of this work, Dobbin *et al.* (2005) published comments on our work. Their remarks related to the design of microarray experiments and notably about the use of dye-swaps. They considered that our recommendations contradicted their previous recommendations Dobbin *et al.* (2003) for designing experiments, where they suggested minimizing or eliminating the use of dye-swap arrays. Following their comments, we precised our point of view in Martin-Magniette *et al.* (2005a) by explaining that our paper focused on the detection, quantification and correction of this gene-specific.

From my point of view, this work is a nice example showing that a methodological work can impact experimental routines. To date, Martin-Magniette *et al.* (2005b) is cited 46 times and the origin of the bias was investigated in Kelley *et al.* (2008) and Margaritis *et al.* (2009). Our paper is often cited to justify the use of dye-swaps. Indeed the existence or not of this bias is important and if it is sizeable, it should be taken into account when experiments or new probes are designed. On the transcriptomic platform of URGV, despite the absence of gene-specific dye bias on the CATMA microarray, it was decided to continue the transcriptome comparisons with dye-swaps to assure quality data.

3.2.3 Bleeding effects in multiple-target microarray

Most of microarray studies are made using one or two dyes labelling which allows the hybridization of one or two samples on the same microarray. In such experiments, the most frequently used dyes are $Cy3$ and $Cy5$. To hybridize simultaneously three or four samples, the dyes *Alexa* 488 and *Alexa* 594 are proposed and Forster *et al.* (2004) evaluated triple-target microarray by comparing results of single-target, dual-target and triple-target microarrays. They concluded that the use of triple-target microarray is valid from an experimental point of view. One year later, Staal *et al.* (2005) investigated the four-target microarray experiments. Their study shows that *Alexa* 594 is best suited as a third dye and that *Alexa* 488 can be applied as a fourth dye on some microarray types.

Introduction of additional dye induced signal bleeding from one dye-labelled sample to another and it is a potential source of bias. Indeed, bleeding artificially increases the signal in other channels of the same spot when the signal is high in one channel. Assume that a gene is highly expressed in condition A and weakly expressed in condition B. The difference between the two conditions is decreased by the bleeding. Therefore bleeding may induce a lowering in the statistical power for detecting differentially expressed genes. Another possibility is that the

bleeding effect induces a difference between two channels for the same gene: assume that a gene is highly expressed in condition A and equally expressed in conditions B and C; if the bleeding between the channel corresponding to condition A and the channel corresponding to condition B is higher than the bleeding between A and C, then a difference between signals B and C will appear, which is a technical artifact.

Forster *et al.* (2004) pointed out that signal bleeding can be considered as negligible between $Cy3$ and $Cy5$ signals, but is important between $Alexa594$ and $Cy3$ signals, so that signal cross-talk can not be neglected. To further investigate this bias, we wrote the signal model to understand the phenomenon and proposed an easy experiment to evaluate and correct the bleeding.

Identification and quantification of the bleeding In order to investigate bleeding, we used single target hybridization microarray experiment produced by the transcriptomic platform of URGV and the single target hybridization data set from Forster *et al.* (2004). From the raw data (not log-transformed) of these datasets, we observed graphically that the bleeding bias $Cy3 \rightarrow Cy5$ is negligible but the bleeding bias $Alexa594 \rightarrow (Cy5, Cy3)$ exists.

To quantify the cross-talks, we explained the blank channels by a linear regression of the hybridized channel:

$$Y_{ij} = \alpha_l + \beta_{lj} Y_{il} + \varepsilon_{ij},$$

where Y_{ij} is the raw intensity of gene i on channel j, β_{lj} the coefficient of bleeding from channel l to channel j, α_l the intercept and ε_{ij} is assumed to be distributed according to a centered Gaussian of variance σ^2. Estimation of the bleeding effect is done using a robust method (R-function *rlm*, Huber (1981)) to decrease the effect of outliers.

The results show that the impact of bleeding on the signal is low: The greater coefficient is between $Cy3$ and $Alexa594$ (0.07). The weakness of the quantitative influence of bleeding is confirmed by the values of the standard error of the signal in the different channels: the values for the empty channels are between 6 and 200 times lower than the corresponding values for hybridized targets. Moreover a good experimental design, where dyes are balanced, is also a solution for cutting down the bleeding bias since the expression differences between two conditions is the mean of the individual measures of this difference taken on each microarray. For example, if only one difference is distorted by the bleeding bias, its influence on the mean difference of expression is divided by the number of terms in the mean, which is equal to the number of microarrays containing the two conditions.

Despite the absence of bleeding observed on the two datasets, it is possible that other dyes or other laser technologies induce a greater bleeding bias. Hence we proposed a procedure to correct the bleeding when its level is high. As the bleeding seems to work on a linear scale, a natural idea is to correct the raw data using the following expression:

$$\widetilde{Y}_{ij} = Y_{ij} - \sum_{l \neq j} \widehat{\beta}_{lj} Yil \tag{3.4}$$

where Y_{ij} is the raw measure of expression of gene i on channel j, $\widetilde{Y}_{ij}$ is the corresponding value corrected for bleeding, and β_{lj} is the coefficient of bleeding from channel l to channel j, estimated from the previous models used to evaluate the bleeding from one-target microarrays.

The proposed correction assumes that the bleeding coefficients do not depend on the intensity of the bleeding channel (this assumption seemed to be valid according to the two analyzed datasets), that the effects of the bleeding from several channels are additive on a linear scale, which is a realistic assumption. Finally, since we used preliminary single-target experiments, it means that the bleeding coefficients do not depend on the microarrays. This latter assumption seems reasonable if the machine-tuning parameters remain constant.

A smarter solution could be to estimate the coefficients β_{lj} by using p-target microarrays but it is not routinely feasible due to the size of the design matrix (more than $2np$, where n denotes

the number of probes). Moreover simulations showed that there are many confounding effects and consequently the estimates of the β_{lja} are not reliable.

3.2.4 Generalization of the lowess procedure for multiple-target microarray

Even if the bleeding bias is a new important bias for multiple-target microarray, the dye bias still remains the most important one. On two-color microarray, the lowess procedure is the most effective method to correct this bias. But its application is not straightforward since the lowess procedure is devised only for correcting the bias between two dyes. A solution proposed by Forster *et al.* (2004) is to applied the lowess procedure sequentially to correct $Cy5/Cy3$, $Cy5/Alexa594$ and $Cy3/Alexa594$. However this procedure does not correct globally the dye bias due to the three dyes. For this reason, we proposed a generalization of the lowess procedure to correct the dye bias in one step on triple-target microarray.

Let $i = 1, \cdots, n$ be the probe index, $j = 1, \cdots p$ the channel index and, y_{ij} the $\log_2$ transformed intensity measure of probe i along the channel j. Let $\overline{Y}_i = \frac{1}{p}\sum_j Y_{ij}$, be the *mean channel* raw data for probe i on the log scale, and $D_{ij} = Y_{ij} - \overline{Y}_i$, the difference between channel j and the *mean channel* for probe i. We model D_{ij} as follows:

$$D_{ij} = f_j(\overline{Y}_i) + E_{ij}$$

and estimate f_j via a lowess. The value of the channel j after normalization is then defined by:

$$\widetilde{Y}_{ij} = Y_{ij} - \widehat{f}_j(\overline{Y}_i) \tag{3.5}$$

Kerr *et al.* (2002) proposed to validate a given normalization method by analyzing the raw and the normalized data through an analysis of variance model. A good normalization method should cut the sum of squares due to technical factors or interactions and should not decrease the sum of squares of the gene-condition interaction. As expected, the normalization reduces all the technical biases and the gene-condition interaction is only slightly decreased.

We measured also the impact of this normalization on the number of differentially expressed genes by analyzing self-experiments, where only one sample is labeled with all the dyes and then hybridized on the same array. In such experiments, no differentially expressed gene is expected. Differential analyses with Varmixt (Delmar *et al.*, 2005) of the triple-self arrays of Forster's experiment and of the URGV dataset give no genes differentially expressed after normalization.

In the framework of the ANR project AgriArray, it was possible to generate the transcriptomes of the same samples with two or three-target microarrays. After differential analysis with Varmixt, the number of differentially expressed genes with three triple-target microarrays is higher than the number of differentially expressed genes with six two-color microarrays. It proves that the proposed normalization for triple-target microarrays does not reduce the true difference between gene expression more than the usual lowess method for two dyes does.

3.3 Model-based clustering for whole-genome analysis

3.3.1 MixThres: Truncated Gaussian Mixtures for an hybridization threshold

This work is the result of a collaboration with Franck Picard

Microarrays are now usually used to investigate a biological process. The principle is to alter the process by a gene mutation or an environment modification and to compare the transcriptome under this constrained processes to identify key genes. There exists another question, more delicate, which concerns the identification of the transcripts in a sample. An answer could be deduced from the transcriptome comparisons. Indeed a gene declared differentially expressed is inevitably transcribed in one condition. Nevertheless the answer is incomplete since a gene could

be not differentially expressed between two conditions because transcribed at the same level in the two samples. Consequently, identification of transcripts in a sample is another question and it requires adapted statistical methods.

In a microarray, there exists a continuum of intensity values among probes and it is difficult to distinguish probes for which there is not enough signal resulting from hybridization with specific target, and not enough identity between the probe and the labeled transcripts to allow hybridization. Existing procedures were usually based on an estimation of a local background per probe from image acquisition softwares and an arbitrary threshold is defined for each individual probe. For example Quackenbush (2002) proposed to declare a probe as being above the background if its red and green intensities are more than two standard deviations above background. These methods are strongly dependent on the estimation of the background, which may be a poor measure of non specific fluorescence Brown *et al.* (2001). Alternatives are proposed by Bilban *et al.* (2002); Stolc *et al.* (2004) based on the knowledge of either positive and negative controls or a population of non-hybridized probes. Nevertheless such knowledge is rarely available. In this work, our aim was to provide an automatic identification of hybridization threshold below which a probe is considered as a non-hybridized probe. First we model the intensity histogram by mixtures of distributions to capture different populations of probes and then we define an hybridization threshold based on the conditional probabilities of the mixture.

Modeling the intensity signal distribution As illustrated on Figure 3.2, we faced two main characteristics of the data: intensity values are bounded and their distribution is positively skewed. Lower and upper boundaries are due to DNA autofluorescence and to saturation respectively. As for skewness, it is due to the high proportion of probes which are not hybridized.

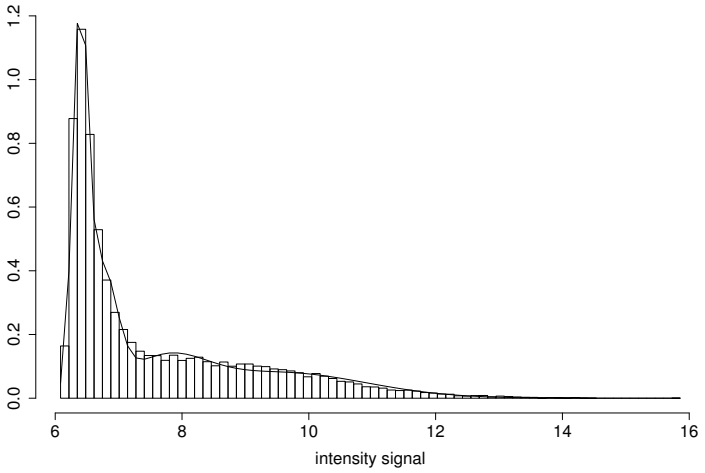

Figure 3.2: Classical intensity histogram

In a preliminary analysis, many usual distributions were tested including non-symmetrical distributions such as Lognormal, Gamma or Weibull distributions, but none systematically fitted the left peak of the empirical distribution. Consequently we introduced truncation parameters $\ell = \min_g(y_g)$ and $u = \max_g(y_g)$, where y_g denotes the $\log_2$ intensity of the probe g, to model this asymmetry and we considered mixtures of truncated Gaussian distributions defined by:

$$g_\ell^u(y; p, \theta) = \sum_{k=1}^{K} p_k g_\ell^u(y; \theta_k),$$

where $g_\ell^u(y; \theta)$ denotes a Gaussian density truncated at ℓ and u, $\theta = (\theta_1, \ldots, \theta_K)$ is the parameter vector describing the truncated Gaussian densities and $p = (p_1, \ldots, p_K)$ with $\sum_{k=1}^{K} p_k = 1$

is the mixing proportion vector. The key element of this model is the introduction of truncation parameters that re-weighted the densities on the finite support $[\ell, u]$. This strategy allows us to model the left peak observed on real intensity histograms. When using truncated Gaussian distributions, the empirical estimators of the mean and of the variance are biased due to truncation. Consequently we added a fixed-point algorithm in the M-step to correct the bias, using the fact that the maximum likelihood estimators are equal to the moment estimators when ℓ and u are known (Johnson *et al.*, 1994).

To fit best the histogram, we considered a collection of mixture models of untruncated, left, right and left-right truncated Gaussian distributions for which the number of components varied between 1 and a maximal number chosen by the user. The best model is then chosen according to the BIC criterion. An ideal situation would be to select a mixture of two components only. However, it appears that fitting intensity histograms with two components leads to poor fit quality and BIC always selects mixtures with four or five components, reflecting better the continuum of intensity values observed in a microarray experiment.

Definition of the hybridization threshold Once the best model is selected using BIC, the components are ordered according to their mean value. Then our aim is to define a hybridization threshold to distinguish non-significant hybridization signal from significant hybridization signals. The only certain information is that the component with the highest mean is composed of hybridized probes, but nothing can be inferred for other components, since there is still an ambiguity between truly hybridized probes with low intensity and non-hybridized probes.

If we consider only probes assigned in the last component as hybridized, it results on a conservative procedure. This is why we proposed the following threshold, $T(\varepsilon)$ above which a probe is declared as being hybridized:

$$T(\varepsilon) = \max \left\{ y_g \mid k_g^* < \widehat{K} \; \exists s \in \{1, \ldots, k_g^* - 1\}, \tau_{sg} \geq \varepsilon \right\},$$

where $\widehat{K}$ denotes the number of components of the selected mixture, τ_{sg} denotes the conditional probability of probe g to belong to the component s.

To define this threshold in practice, it is equivalent to rank intensity values by descending order. Then sequentially for each probe, the component for which its conditional probability is the highest is determined. If this component differs from the component with the highest mean, we look at the conditional probability of belonging to each component with lower mean. If none of the conditional probabilities is greater than ε, then the next probe is analyzed. Otherwise, if one of the conditional probabilities is greater than ε, then the threshold $T(\varepsilon)$ is defined as the intensity value of this probe. Our method is implemented in an R package called MixThres with default value of ε equal to $\varepsilon = 10^{-4}$.

Performances of MixThres We validated our method using four biological samples hybridized on a tiling microarray of the whole chromosome 4 of *Arabidopsis thaliana*. The specificity of this array is that probes cover genic as well as intergenic regions and when hybridizing labeled mRNAs only, we expect that probes corresponding to intergenic regions should not hybridize. For each biological sample and each probe of the tiling array, we calculated a hybridization index which equals 1 if the signal intensity is higher than the hybridization threshold and 0 otherwise. We get a hybridization score between 0 and 4, defined as the sum of the hybridization index over the four samples for each probe. Out of the 21602 probes, 4681 are declared hybridized four times (score=4) and 13681 are never declared hybridized (score=0), thus 85 % of the results are coherent between the four biological samples. Moreover, the threshold as well as the percentage of probes declared hybridized vary between the four biological samples reasonably. This shows that our method provides reproducible results.

By definition the specificity is the probability to declare a probe not hybridized rightly. To estimate it, we focus on the set of 3701 probes which cover intergenic regions without any

proof of transcription or cross-hybridization. Among them, 49 probes are declared hybridized. Consequently the specificity is estimated at least 98.7 %. Similar estimate of the specificity is found when the same procedure is applied to another experiment.

We also estimated the precision which is the proportion of probes declared rightly hybridized amongst all probes declared hybridized. In Aubourg *et al.* (2007), we analyzed the 522 hybridized samples with MixThres and for all probes without any known evidence of transcription declared hybridized at least once, an experimental validation was performed to validate this result. Among the 465 new genes found hybridized at least once in the 522 hybridized samples, the hybridization evidence was confirmed by other experimental approaches for 88%. The biological results of this project are detailed in Section 4.2.

Some comments We provided an automatic definition of a hybridization threshold. From the modeling point of view, considering the truncation leads to a better fit of the left peak of intensity histograms. On the studied datasets we observe that the first smallest value of BIC is associated with a truncated Gaussian mixture model and the second smallest value with the untruncated Gaussian mixture with the same number of components. Interestingly both hybridization thresholds derived from these two models are equal. This indicates that the threshold value is well-defined and does not depend on the truncation parameters.

The hybridization threshold depends on a parameter ε, which has been set at 10^{-4} in this study. Consequently the specificity and sensitivity depend on ε since a greater value will lead to a decrease (increase) in specificity (sensitivity). In order to assess the role of ε, we have tested several values from 10^{-3} and 10^{-6} and estimations of the specificity and the precision are the same.

We submitted MixThres several times in bioinformatics journal but it was always rejected. Even, MixThres was a very useful method, applied in three different biological projects published in biological journals. These projects are described in Chapter 4.

3.3.2 MultiChIPmix: Linear regression mixture for epigenomic data

This work corresponds to the internship of Caroline Bérard, co-supervised with Tristan Mary-Huard

Background Chromatin immunoprecipitation (ChIP) is a well-established procedure used to investigate proteins associated with DNA. ChIP on chip involves analysis of DNA recovered from ChIP experiments by hybridization to microarray. In a two-color ChIP-chip experiment, two samples are compared: DNA fragments crosslinked to a protein of interest (IP), and genomic DNA (Input). The two samples are differentially labeled and then co-hybridized on a single array (see Figure 3.3). The goal is then to identify actual binding targets of the protein of interest, i.e. probes whose IP signal is significantly larger than the Input signal. Two strategies are widely applied for the detection of enriched DNA regions. The first strategy takes advantage of the spatial structure of the data: Since probes are positioned all along the genome, if one region is enriched, several adjacent probes are expected to obtain high ratio measurements, resulting in a "peak" of intensity. To detect these peaks, sliding windows is proposed by Cawley *et al.* (2004); Keles (2007) and Hidden Markov Models by Ji and Wong (2005); Li *et al.* (2005). Alternatively, the second strategy considers that the whole population of probes can be divided into two components: the population of IP-enriched genomic fragments and the population of genomic DNA that is not IP enriched. In this framework, different statistical methods are proposed to distinguish between the two populations by considering the distribution of the ratios (or their associated rank). If a non negligible proportion of the fragments are enriched, then the logratio distribution is bimodal and the highest mode corresponds to the enriched population. Buck and Lieb (2004) proposed to declare a probe enriched when its ratio exceeds a selected cutoff, fixed according to the logratio distribution.

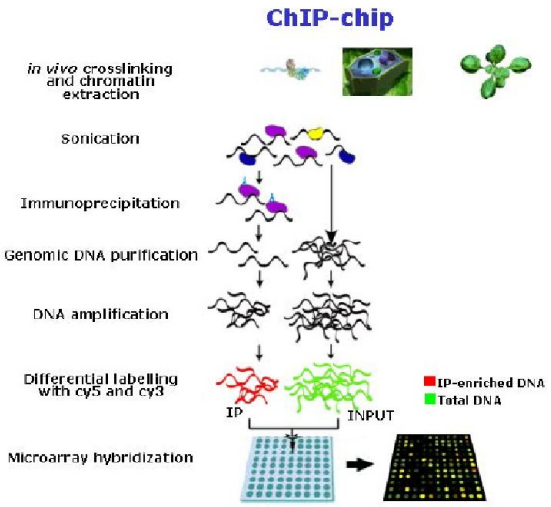

Figure 3.3: Principle of a chIP-chip experiment

Questioning the logratio analysis Importantly, both strategies assume that the logratio measurement is a pertinent statistical quantity to assess the probe status (enriched or not). We discovered that this assumption is correct if the distribution of the ratio mostly depends on the status of the probe. Figure 3.4 (left) shows the ideal situation described in Buck and Lieb (2004), where the logratio distribution is bimodal. But in many applications, the logratio distribution is closer to Figure 3.4 (right), and the performance of logratio-based methods may be poor. From this observation, we argued that it is worth working directly with the two measurements of each probe (Input and IP) rather than on the logratio.

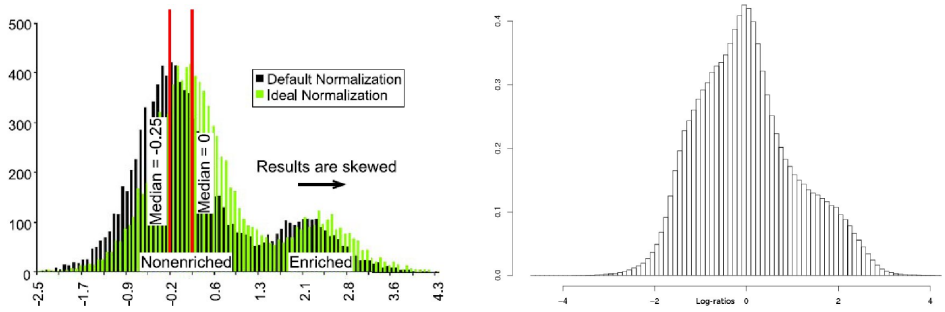

Figure 3.4: **Left:** Ideal logratio distribution with two distinct peaks. **Right:** Logratio distribution on a real data (NimbleGen array).

We also observed that the relationship between the two measurements is almost linear (Figure 3.5). Working on logratio amounted to stating that the slope of the linear relationship is the same whatever the status of the probe. In many cases the slopes are different and thanks to simulated data we showed that even a slight difference between the two slopes may turn the distribution of the logratios into unimodal rather than bimodal, as observed in Figure 3.6.

Modeling the joint distribution of the IP and Input intensities In this context, we proposed a new model of the distribution of the IP conditionally to the Input intensity to detect enriched probes in a chIP-chip experiment. In Martin-Magniette *et al.* (2008a), we proposed a mixture model of two linear regressions to characterize the IP-Input relationship of one experiment. In Bérard *et al.* (2013), we generalized the model to analyze simultaneously several independent replicates and to take into account possible spatial dependencies between adjacent

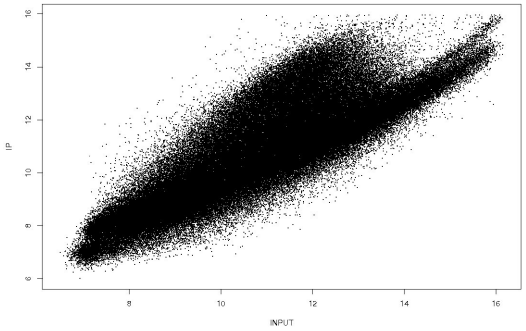

Figure 3.5: Associated plot of IP versus Input

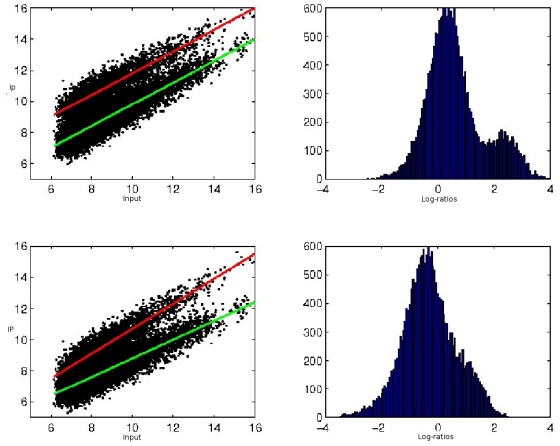

Figure 3.6: Simulated data. **Top:** Two populations with linear relationship and equal slopes. The corresponding logratio histogram is bimodal. **Bottom:** Two populations with linear relationship but different slopes. The corresponding logratio histogram is unimodal.

probes.

Formally, let (x_{ir}, y_{ir}) be the pair of log-Input and log-IP intensities of probe i measured in replicate r of a ChIP-chip experiment and denote the (unknown) status of the probe through the latent variable Z_i equal to 1 if the probe is enriched and 0 if it is normal. It means that

$$f(y_{i1}, \ldots, y_{iR} | x_{i1}, \ldots, x_{iR}, z_i) = \prod_{r=1}^{R} \mathcal{N}(a_{z_i r} + b_{z_i r} x_{ir},\ \sigma_r^2)$$

with specific mean $a_{z_i r} + b_{z_i r} x_{ir}$ and variance σ_r^2 for each replicate $r \in \{1, \ldots, R\}$

If the ChIP-chip experiment is performed on a microarray where a small number of probes cover promoter regions, then all probes are assumed to be independent and the distribution of the latent variable Z_i is a Bernoulli of parameter π. The model is hence a mixture of linear regressions called MultichIPmix. Parameters (proportion, intercepts, slopes and variances) are estimated using the EM algorithm. For the initial values, the proportion and the variances are set to default values and the slope and intercept of each replicate are derived from the first axis of the Principal Component Analysis to reduce the sensibility of the EM algorithm to the initial values.

If the ChIP-chip experiment is performed on a microarray where probes cover the whole genome, then several probes cover a given biological unit and spatial dependence exists between adjacent probes. We proposed to take it into account by modeling the distribution of the latent variable $\{Z_i\}$ by a Markov chain. The model is called MultichIPmixHMM. Parameters of the

HMM are estimated using the Baum-Welch algorithm (Rabiner, 1989). For the initialization, for each replicate, slope and intercept are derived from the first axis of a Principal Component Analysis by considering a mixture model. All initial transition parameters are set to 0.5.

A binary latent variable is adapted if the protein under study has some targets. When the protein has no target, all probes belong to the normal class. In this case a simple linear regression is sufficient to fit the data. For this reason, we recommend to fit the two models (one or two classes) and to select the best model according to the BIC criterion.

Classification rules Whatever the modeling of the latent variable, for each probe, we get a conditional probability to be enriched and a probe is declared enriched if its enriched conditional probability is higher than $1 - \alpha$, where α is chosen by the user. This strategy controls the proportion of misclassification in mixture models (Mary-Huard *et al.*, 2013). This is during this work on ChIP-chip analysis that we defined a classification rule to control the false positive risk presented in Section 2.3.3.

Performances In Martin-Magniette *et al.* (2008a), we presented various applications each dedicated to one specific biological question (histone modification and DNA methylation on different organisms) to be convinced and convince the biologists that our model is convenient for any microarray whatever its density (array size from thousands to hundreds of thousands of probes) and for any organisms.

We re-analyzed the dataset of Weber *et al.* (2007) studying the DNA methylation of promoter regions of the human genome. Each promoter region is covered by 15 probes and is classified into a category according to its CpG rate. We focused on the analysis of the class ICP (intermediate CpG promoter) composed of 2056 promoter regions. We applied MultiChIPmix to these data without averaging the 15 values per promoter region. The estimated regression slopes are $\widehat{b}_0 = 0.613$ for the normal class and $\widehat{b}_1 = 1.162$ for the enriched one, which shows that the Input-Ip relations substantially differ between the two status. At the level $\alpha = 0.01$, a total of 1706 promoter regions have at least 1 probe enriched. Except for one region, all the promoter regions of the Weber's list have at least 1 enriched probes, and 403 have 5 or more enriched probes. Besides, MultiChIPmix identified 38 promoter regions with 9 probes or more classified as enriched that are not detected in Weber *et al.* (2007).

The other example are histone modification of *Arabidopsis thaliana* and the collaboration with the team of Vincent Colot was very helpful since they provided data of good quality and various samples were hybridized on different supports. I focus here only on the H3 tri-methylated at lysine 27 (H3K27me3) histone mark.

First the study was done using a custom genomic tiling array of chromosome 4. These data are first analyzed with MixThres and published in Turck *et al.* (2007). We re-analyzed the data with MultiChIPmix. The tiles classified as enriched at risk $\alpha = 0.01$ include all the tiles found by Turck *et al.* (2007) plus 2346 others: 1404 tiles extend the genomic region already found and 942 tiles form 62 new genomic regions. The difference between the two slopes (0.907 and 1.167) enables us to better discriminate the two classes for high Input intensities and this may explain the higher number of enriched probes detected by MultiChIPmix.

More recently in Bérard *et al.* (2013), we demonstrated improved performance of MultiChIPmixHMM compared to MultiChIPmix. We used data generated from a two-color NimbleGen microarray where each chromosome are covered by about 200000 (Roudier *et al.*, 2011). We applied both methods and identified probes enriched in H3K27me3 using a cutoff of $\alpha = 0.01$. For MultiChIPmix, both replicates are analyzed separately and only probes declared as enriched in both replicates are finally considered as enriched. MultiChIPmix and MultiChIPmixHMM agree for more than 90% of the probes. All enriched probes identified by MultiChIPmix are included in the set of enriched probes identified by MultiChIPmixHMM, which identified $7,940$ additional probes. From a general point of view enriched region are covered by more enriched probes in the MultiChIPmixHMM analysis. We also considered 311 H3K27me3 target genes identified by

independent prior studies by Turck *et al.* (2007) and by Zhang *et al.* (2007). Among the 311 genes, 298 are commonly identified by both methods and 11 are only identified by MultiChIPmixHMM. Consequently, we get a better modeling of the data by considering spatial dependence between adjacent probes.

Biological results obtained by the application of MultiChIPmix in different biological projects are described in Section 4.3.2.

3.3.3 TAHMMAnnot: Heterogeneous HMM for data of tiling array

This work corresponds to models developed by Caroline Bérard during her PhD.

Background To compare two samples from microarray data, most methods rely on the log-ratios. But as we showed in Martin-Magniette *et al.* (2008a), this can mask the multimodality of the data and it is worth working directly with the two intensities of each probe. MutiChIPmixHMM proposed a model describing a sample as a function of the second one. This situation is relevant to detect interactions protein-DNA from ChIP-chip data, but this situation is rare and in most cases, no sample can be considered as a reference.

Comparing two samples requires distinguishing four different biologically interpretable groups of probes: a group with similar behavior in both samples, a group with higher intensity in the first sample than in the second sample, a symmetric group with higher intensity in the second sample and a last group with low intensity in both samples which can be viewed as noise, corresponding to the non-transcribed regions (Figure 3.7). In TAHMMAnnot, we proposed a 4-state heterogeneous hidden Markov model with bidimensional Gaussian emission densities to gather all the available information (the two hybridization signals of the probe, its position along the genome and its structural annotation) to study the difference between two samples. In the following, I describe the different models proposed and show how the modeling improves the data fitting.

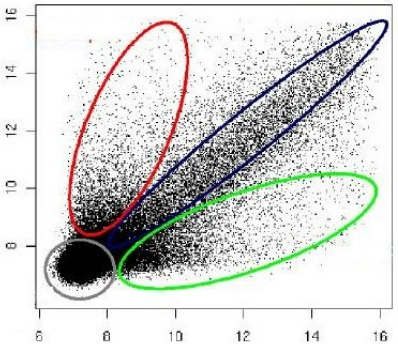

Figure 3.7: Schematic explanation of the 4 groups to consider when comparing two samples. In grey, the noise group corresponding to not hybridized probes. In blue, the identical group, in which probes have the same intensity in both samples. In red, the over-expressed group in which probes have a higher signal in sample of the y-axis and in green the under-expressed group in which probes have a higher signal in sample of the x-axis.

Models for the latent variable Z For probe t, we denote $Y_t = (Y_{t1}, Y_{t2})$ the log-intensities for both samples, $C_t \in \{1, ...P\}$ the annotation category and $Z_t \in \{1, ...K\}$ the unknown status. We proposed several model for the latent variables $\{Z_t\}$: if there is a dependence between adjacent probes, $\{Z_t\}$ are assumed to be a first order Markov chain with heterogeneous transition π^p depending on the annotation category:

$$P(Z_t = l | Z_{t-1} = k, C_t = p) = \pi^p_{kl}$$

If there is no spatial dependence, the $\{Z_t\}$ are independent and distributed according to a multinomial of parameter π_k^p corresponding to the proportion of probes from group k in annotation category p. If there is no annotation and no spatial dependence, the model comes down to a mixture model with four components.

For the conditional distribution of the $\{Y_t\}$, we proposed two models based on Gaussian distribution. Both are presented and then discussed. In the following, $K = 4$ groups are considered, Groups 1 and 2 will refer respectively to 'noise' and 'identical' probes, whereas Groups 3 and 4 will refer to differentially hybridized probes.

Model 1: constrained Gaussian as emission distribution In this first model $\{Y_t\}$ are assumed to be independent conditionally to the $\{Z_t\}$ and distributed according to a Gaussian distribution, whose parameters depending on the probe status but not on the annotation category:

$$(Y_t|Z_t = k) \sim \mathcal{N}(\mu_k, \Sigma_k),$$

where μ_k and Σ_K are the mean and the variance matrix of the k-th emission distribution. To get interpretable groups, we modeled the variance matrix using its eigenvalue decomposition (Banfield and Raftery, 1993)

$$\Sigma_k = D_k \Lambda_k D_k',$$

where the matrix Λ_k describes both the volume and the shape of the ellipse associated with the Gaussian distribution and the matrix D_k describes its orientation. By definition Groups 1 and 2 should have the same orientation (see Figure 3.7), which implies that $D_1 = D_2$. Furthermore we assumed that the dispersion around the main axis is similar in all groups, which amounts to fixing the second eigenvalue of Σ_k for all groups. This can be summarized as

$$\begin{aligned} D_1 &= D_2; \\ \Lambda_k &= \begin{pmatrix} u_{1k} & 0 \\ 0 & u_2 \end{pmatrix}, \qquad \text{with } u_{1k} > u_2, \text{ for } k = 1, ..., 4. \end{aligned}$$

The parameters $\{\pi^p\}$, $\{\mu_k\}$, $\{D_k\}$, $\{u_{1k}\}$ and u_2 are then estimated using an adapted EM algorithm.

Model 2: Gaussian mixture as emission distribution To make the emission distributions more flexible than in the previous setting, we assumed that each emission distribution is itself a mixture of L_k parametric distributions. Working with mixtures of bidimensional Gaussian are not easily tractable, therefore we projected the data on three axes Δ_2, Δ_3 and Δ_4, concurrent at the barycenter of the noise group and corresponding respectively to the main axis of the ellipse representing the three groups. It leads to to work with unidimensional mixtures and force the Gaussian components of the k-th cluster to be colinear along the axis Δ_k (see Figure 3.8).

In this second model,

- The noise group is a special group, considered as circular and modeled by a spherical Gaussian.
- For each of the three other groups, let (U_{tk}, V_{tk}) be the coordinates of (Y_{1t}, Y_{2t}) in the orthonormal basis $(\Delta_k, \Delta_k^\perp)$. An unidimensional Gaussian mixture along each axis Δ_k and a unique distribution for all components along $\Delta_k^\perp$ are considered:

  $$(V_{tk}|Z_t = k) \sim \mathcal{N}(0, \sigma^2) \quad \text{and} \quad (U_{tk}|Z_t = k) \sim \psi_k \ ,$$

  where

  $$\psi_k = \sum_{\ell=1}^{L_k} \eta_{k\ell} \phi(.; \mu_{k\ell}, \sigma_{k\ell}^2) \ ,$$

  with $\eta_{k\ell}$ is the proportion of the ℓ-th component for the group k and ϕ is the density distribution of a Gaussian of mean $\mu_{k\ell}$ and variance $\sigma_{k\ell}^2$.

The parameters are then estimated using an adapted EM algorithm equivalent to the one presented in Section 2.2.2.

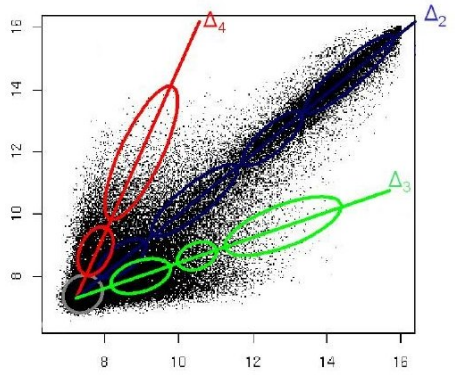

Figure 3.8: Schematic representation of Gaussian mixtures in each cluster, along the three axes.

Comparison of the latent variable modeling Considering constrained Gaussian distribution to model each emission distribution, we evaluated the impact of the modeling of the latent variable distribution. It is based on a real comparison of leaf and seed 10 days after pollination of the plant *Arabidopsis thaliana*. Transcriptomic data are produced with a NimbleGen array with about 200 000 probes per chromosome. For the annotation, three categories are considered to distinguish intergenic, intron and exon probes.

Table 3.2 presents the fit of the four models for chromosome 4. We remark that including annotation information leads to a real improvement in terms of likelihood, both in the mixture and HMM context. Moreover BIC an ICL criteria both suggested that all available information should be taken into account, since the HMM with annotation information is preferred.

Table 3.2: Impact of the latent variable distribution in a comparison of leaf and seed transcriptome samples.

	Mixture	HMM	Mixture + annotation	HMM + annotation
number of parameters	19	31	25	61
-2 log-likelihood	406249	371309	373283	356617
BIC	406469	371668	373573	357323
ICL	436197	412706	399986	398272

Caroline Bérard investigated the behavior of the four models during her PhD. She observed that an HMM tends to smooth the classification. As for example, when an isolated probe in an intergenic region has abnormally high intensities, the HMM tends to declare it not differentially expressed, whereas the mixture could declare it differentially expressed. Nevertheless for intronic probe between two exonic probes, the smoothing of the classification is a drawback (see Figure 3.9) and it can be offset by including the annotation in the modeling. In general, when the latent variable distribution takes into account the annotation and the dependence between adjacent probes, it leads to more homogeneous regions.

Constrained Gaussian or Gaussian mixture for the emission distribution Based on the comparison of leaf and seed 10 days after pollination of the plant *Arabidopsis thaliana*, we compared Model 1 including 3 annotation categories and Model 2. In this second model, the annotation is not taken into account and for the emission distribution, the component number of the mixtures are equal. Figure 3.10 represents the data distribution and the adjustment of the estimated density for each group by a Gaussian mixture. According to BIC, the best number of

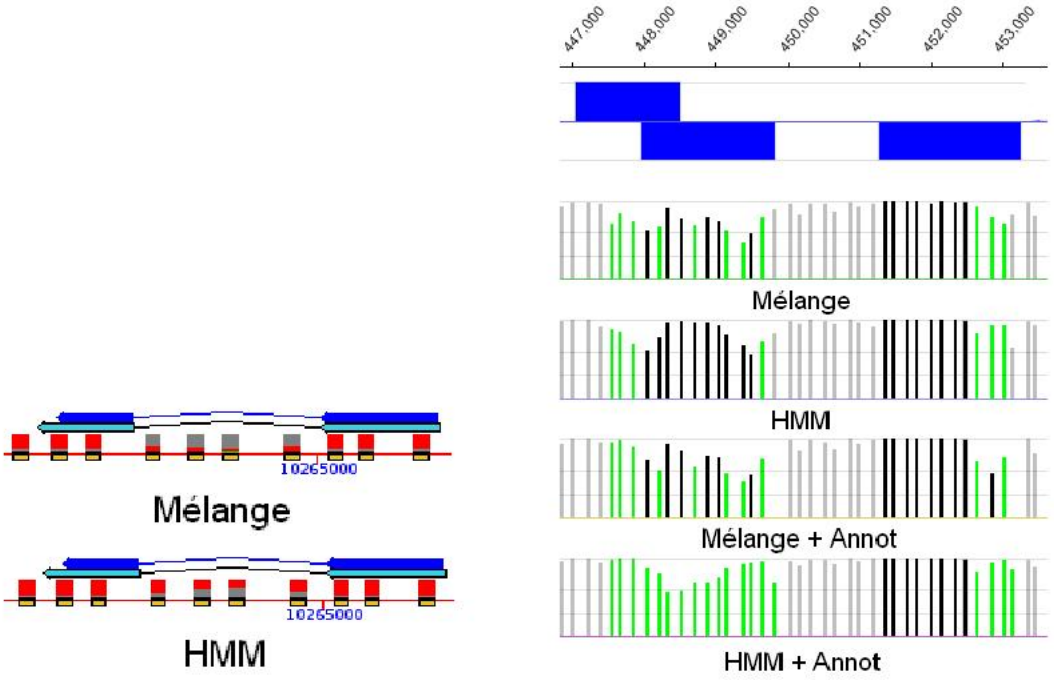

Figure 3.9: **Left:** Example of a gene with two exons (blue boxes) and one intron (line between the two boxes) covered by several probes. The color of each probe is proportional to its conditional probabilities (red= over-expressed, gray= no hybridized). **Right:** example of a region, blue boxes correspond to genes. The classification of the probes for each model is given below, the bar is proportional to the highest conditional probability and its color indicates the group

Gaussian is four (BIC=362142) but between the two modelings, Model 1 is better by taking the annotation into account (BIC=357323). Whereas if the annotation is neglected (BIC=371668), the mixture of Gaussian as emission distribution is selected. To go further, we compared the classification results. Between the two modelings, there is 6% of differences in terms of probe classification in this comparison. Without a bioinformatic analysis and/or a biological validation, it is thus difficult to conclude.

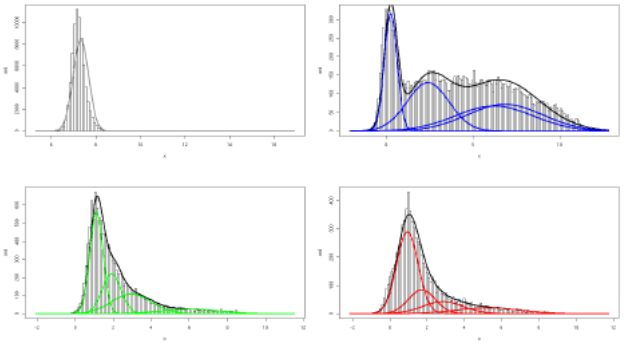

Figure 3.10: Distribution of the data and adjustment of the estimated density for each group (noise in grey, identical in blue, under-expressed in green and over-expressed in red). For the identical, under-expressed and over-expressed groups, the data are projected onto the main axis of the corresponding group.

Recently, I compared two transcriptomes with fewer differences that can exist between transcriptome of leaves and seeds. It is a comparison between a mutant and a wild-type of *Arabidopsis thaliana*. The mutant is expected to accumulate transcripts in some intergenic regions. Moreover, it is expected that differentially expressed probes are only over-expressed in the mutant. After discussion with the biologists, it was decided to not include the annotation in the model to have more chance to detect the differentially expressed intergenic regions. I tried to analyze the data with Model 1 and it was impossible because the EM algorithm did not converge or because groups were not biologically interpretable. The analysis with Model 2 allowing Gaussian mixture as emission distribution was difficult because the EM initialization was delicate but we managed to analyze the data and the results are validated by another technology.

Detection of new transcripts Another major challenge is the detection of new transcripts and transcriptomic data can reveal new transcriptional units that had escaped the conventional annotation methods because of their original structure (small length, antisense, RNA genes, etc.). With the distinction between the noise group and the identical group, TAHMMAnnot without annotation can be relevant to find new transcripts. As for an example, in the transcriptomes of leaves and seeds, a lot of regions with expressed probes are found in intergenic regions: 1328 small regions with 2 or 3 consecutive probes, 185 regions with 4 or 5 consecutive probes and, 90 regions with more than 5 consecutive probes (including 25 regions with more than 10 consecutive probes). For the 90 regions with more than 5 consecutive probes, we checked with other annotation information such as Expressed Sequence TAG (EST) or genes predicted by the Eugene software (Schiex *et al.*, 2001) which are not yet in the official annotation. We found 39 regions matching with annotation like small RNA, rRNA, tRNA, including 12 regions corresponding to a coding sequence defined in Eugene and 10 corresponding to transcriptional units recently annotated due to the presence of EST. Moreover the obtained results showed many transcriptions in the introns in 5'UTR (40% of intronic probes declared expressed), which is consistent with a recent article of Cenik *et al.* (2010) assuming a functional role of 5'UTR short introns.

3.4 Conclusions and perspectives

For developing statistical methods for a dedicated application domain, it requires to have knowledge on this domain and to collaborate with scientists of another disciplinary field. Concerning my works, they are done in closed collaboration with the biologists. Thanks to various collaborations, numerous discussions with biologists and the labmeeting of the URGV, I acquired a general culture in genomics. It allows me to easily interact on genomic projects to identify the questions and understand the various points of view.

All high-throughput technologies are difficult to control and despite efforts to establish rigorous protocols, some technical biases occur and I think that it is of crucial importance to study them so that users understand their impact on the results. Their identification is never easy and provides only the half of the answer since it is important to propose a correction or an arrangement so that biologists pursue their work. The statistical modeling of the signal is the key ingredient in the normalization. It allows the statistician to quantify the bias and its consequences if it is neglected. Data normalization always takes a long time and the normalization methods are difficult to publish. Nevertheless a good knowledge of the technology guarantees its exploitation at the best.

Developing model-based clustering for analyzing genomic data come from discussions on the genome-wide activity and the need for some biologists to look at the transcriptome data as the response of an organism rather than the response of numerous genes. It provides a statistically rigorous framework and has the advantage of providing straightforward procedures for parameter estimation as well as a conditional probability for each observation of belonging to each cluster. Moreover many model selection criteria are available to choose the number of groups without

a priori knowledge. Consequently the main difficulties are to well understand the biological question and what kind of information is expected. For the three models developed, I had many animated discussions. The biologists are often more enthusiasm than I. They often remark the gain of a modeling to their questions whereas I focus on the situations that are not well modeled. All this work requires also a substantial work of programming to automate the analysis, in a first part to evaluate the robustness of the modeling on different datasets and in a second part to transform the method into a tool, usable for the bioinformatic community and also for the biologists willing to analyze themselves their data. I usually write R functions and I am trying now to integrate them in R packages.

Molecular biology is mainly driven by the technology advances but the technologies change and the questions remain. As for an example, the main biological questions when biologists investigate transcriptome are which genes are transcribed and which genes are altered in their expression when different conditions are considered. TAHMMAnnot tries to answer with only one model to both questions by taking all available information into account: the two intensities of each probe, the dependence between adjacent probes and the structural annotation. Unfortunately, HTS technology arrived too early (or we were too slow) and tiling array are no more proposed to explore the transcriptome of a genome. I still remain optimistic because the modeling is a technique to take distance to the application.

Concerning High Throughput Sequencing, research of the appropriate statistical methods for the analysis of digital gene expression has flourished, primarily in the context of normalization and differential analysis. In 2010, with Andrea Rau, Cathy Maugis-Rabusseau and, Gilles Celeux, we decided to go further in the analysis by directly focusing on the question of clustering digital gene expression profiles as a manner to discover groups of co-expressed genes. The first months were difficult because the technology was still young and many biologists were less interested in the modeling questions than in the biological results quickly obtained thanks to the high sensitivity of the HTS. Fortunately the transcriptomic platform of URGV obtained the funding to start a technological project aiming at comparing array technology and HTS technology. They produced the transcriptome of various samples with the well-controlled microarray CATMA and the new sequencing technology. After numerous discussions, we better understand the data and reformulated the biological questions according to the specificity of HTS data. Actually, we are developing three mixtures for modeling RNAseq data. The first one is a mixture of Poisson loglinear models to cluster RNAseq data. The first challenge was to define a pertinent mean modeling of the Poisson distribution used in a mixture to cluster co-expressed genes (Rau *et al.*, 2011). The second model is a constrained version of the first one to identify genes differentially expressed between two conditions. The last one concerns a new question that can be addressed with the HTS technology. I think that it is possible to describe a transcriptome as a function of another one. To answer to this question, we developed a generalized linear Poisson mixture model and proposed solutions to solve some initialization problems of the EM algorithm due to a large number of components of the mixture. It was done by Panagiotis Papastamoulis during a one-year post-doctoral position and he developed an adaptive initialization scheme to avoid local maxima and to get efficient estimation of the parameters (Papastamoulis *et al.*, 2013).

Chapter 4

Contribution to genomic projects

I am deeply involved in genomic projects and this chapter is devoted to my contributions to these projects. The aim of this chapter is to show the added-value of the statistics for the genomic projects. The first section describes my activity on the transcriptome platform of URGV. Additionally to this collaboration, I also collaborate on some specific projects, in particular dealing with the control mechanisms of gene expression. This is the topic of the second section of this chapter.

4.1 Creation of a genomic resource

4.1.1 Background

The main goal of the transcriptomic platform of URGV is to offer an expertise on development and analysis of transcriptomic data in model and non-model plant species to the plant science community. A large panel of transcriptomic technologies (from regular DNA and tiling microarrays to deep sequencing) is offered to be able to propose the technology best suited to the biological questions and plant material provided by the users. The involvement of the transcriptomic platform includes 4 main steps:

1. helping laboratories in developing their experimental design
2. performing all the experiment (library production, sequencing, hybridizations)
3. undertaking the statistical analysis
4. completing data submission in both CATdb (dedicated to URGV transcriptomic experiments) and GEO (NCBI international database) databases.

For Arabidopsis transcriptome analyzes, the transcriptomic platform developed a transcript microarray named CATMA[1], in the European program framework initiated in 2000. This array was initially based on Genes Specific Tags (GST) which are short sequences specific for each gene of Arabidopsis. Several versions of the CATMA array were designed, each new version completing the previous version. The last version, CATMA v6.2, is now based on an oligo array technology where amplified GST are replaced with long primers synthesized on the array. This last version is a very complete array containing probes for all the Arabidopsis genes, including mitochondrial and chloroplast genes, repeat elements, miRNA/MIR, and other non coding RNAs (rRNA, tRNA, snRNA, soRNA).

For crop transcriptome analyzes, the transcriptomic platform proposes an Affymetrix platform where protocols were optimized for plant samples. They also propose custom made oligo microarrays by designing probes from an existing genome, a set of Unigenes or from the deep sequencing analysis of cDNA pools.

[1] Complete Arabidopsis Transcriptome MicroArray

More recently, the platform has also developed a High-Throughput Sequencing (HTS) pipeline for a quantitative transcriptome analysis of model and cultivated plants.

The platform is open for 10 years and Figure 4.1 shows its output in terms of technology. At the beginning, the platform has offered only Arabidopsis transcriptome analyzes. In 2006, the Affymetrix system opened to propose crop transcriptome analyzes. In 2007, we started an ANR project to develop a tiling array of Arabidopsis. In 2009, the technology was enough controlled to be added in the services of the transcriptomic platform. Since 2011, due to important improvements of the high-throughput technologies, the platform manages six different technologies. At each time a new technology is proposed, I have to associate a statistical analysis. This is described in Section 4.1.3

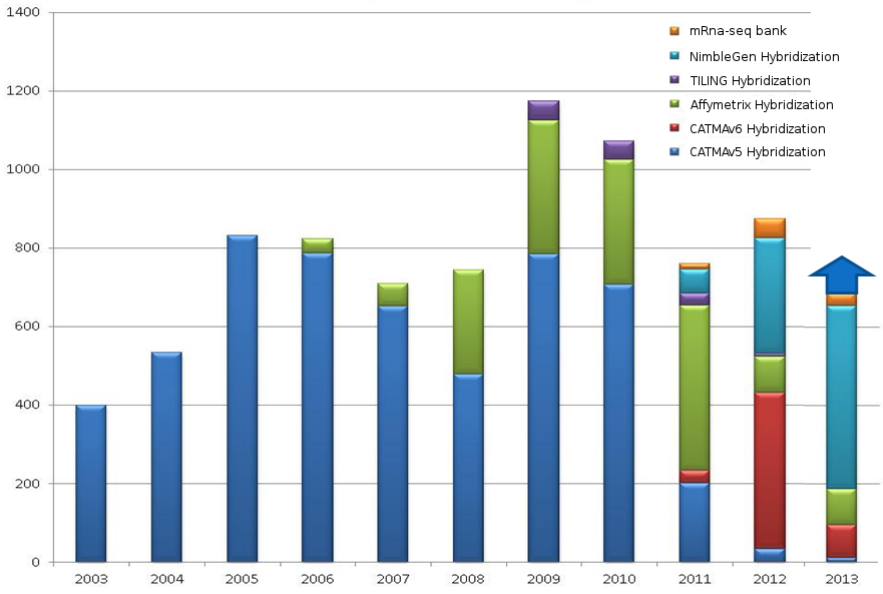

Figure 4.1: Evolution of the technologies proposed by the transcriptomic platform of URGV

Since its opening in 2003, all members of the platform pays great attention to the quality of services offered. The platform obtained the IBISA label and was defined as strategic platform by INRA in 2008. In 2012, the platform obtained the ISO9001 certification including a statistical and bioinformatic process.

4.1.2 CATdb: a database for the transcriptomic platform of URGV

All the data are managed in a dedicated database CATdb developed by the team "Bioinformatics for predictive genomics" of URGV and published in Gagnot *et al.* (2008). CATdb is a free web service available at http://urgv.evry.inra.fr/CATdb. It allows the community of biologists an easy access to the data. The complete description of the experiments is submitted via a private web interface that helps to respect the MIAME instructions. Data stored in CATdb are also available either at GEO or at ArrayExpress.

The CATdb schema and objects are based on those used in the ArrayExpress database (Rustici *et al.*, 2013). The main differences are

- the possibility to manage a supplementary step with the pooling of samples or extracts
- the systematic addition of a figure describing the design of an experiment in standardized format. A double arrow corresponds to a dye-swap, the number of biological replicates is indicated by the number of double-arrows. The green box corresponds to the sample labeled with the green dye on the first slide of a dye-swap. This figure is a very useful

general overview of the experimental design. Figure 4.2 (top) gives the experimental design of the project AF30_Starch_circadian_rythm.

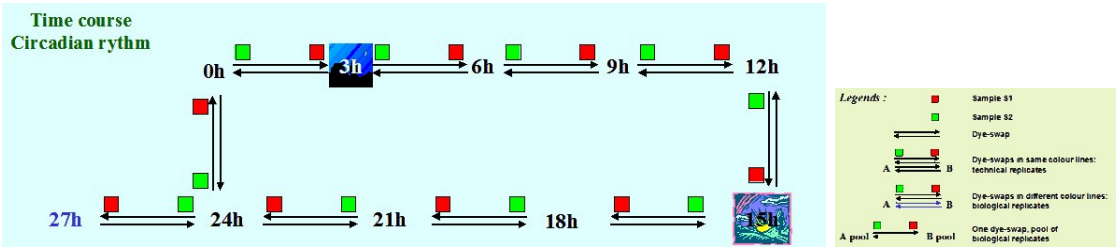

Figure 4.2: The experimental design describing all the hybridizations on the project AF30_Starch_circadian_rythm. Comparisons are done in dye-swap on only one biological replicate.

- the storage of the normalized data and of the results of a differential analysis performed on each dye-swap. In the other databases, only raw data are provided. In CATdb, we decided to provide the normalized data to allow anybody to start a new study without being obliged to normalize again the data. The results of the differential analysis consist in both normalized intensities, the log-ratio and the Bonferonni P-values. To help the readers, a results are encoded using colors to facilitate the result interpretation. Figure 4.3 illustrates the results of a query of CATdb for the project named AF30_Starch_circadian_rythm.

191 probes differentially expressed, sorted for the dye-swap T12 / T9 page 1 / 4

organs:		leaf/leaf				leaf/leaf				leaf/leaf				leaf/leaf			
dye-swap name:		T12 / T9				T15 / T12				T18 / T15				T21 / T18			
GST ID	GENE ID	I S1	I S2	R	P-VAL	I S1	I S2	R	P-VAL	I S1	I S2	R	P-VAL	I S1	I S2	R	P-VAL
CATMA3A01355	AT3G02380	11.58	8.26	3.325	0E+00	12.71	10.96	1.751	0E+00	10.25	12.23	-1.985	0E+00	8.69	10.45	-1.760	0E+00
CATMA1A00045	AT1G01060	12.52	9.72	2.804	0E+00	12.76	11.88	0.877	6E-04	10.13	12.82	-2.698	0E+00	6.78	9.67	-2.888	0E+00
CATMA2A45275	AT2G46830	9.28	6.96	2.322	0E+00	9.66	9.11	0.547	1E+00	8.62	9.40	-0.778	1E-03	6.30	7.94	-1.644	0E+00
CATMA2A39600	AT2G41250	10.41	8.15	2.263	0E+00	11.65	9.96	1.692	0E+00	9.89	11.80	-1.908	0E+00	7.82	9.66	-1.840	0E+00
CATMA5A17020	AT5G18670	12.02	9.89	2.131	0E+00	12.49	11.79	0.700	2E-01	9.40	12.44	-3.041	0E+00	6.99	8.60	-1.610	0E+00
CATMA4A35650	AT4G33870	8.99	6.88	2.105	0E+00	12.15	8.76	3.389	0E+00	12.03	11.54	0.490	1E+00	8.57	10.00	-1.430	0E+00
CATMA5A58030	AT5G62430	9.10	7.00	2.099	0E+00	9.23	8.80	0.430	1E+00	8.87	9.27	-0.396	1E+00	7.43	8.21	-0.778	1E-04
CATMA1A31190	AT1G32900	10.16	8.17	1.993	0E+00	11.09	9.52	1.571	0E+00	10.24	10.93	-0.693	3E-02	7.68	9.93	-2.252	0E+00
CATMA2A29210	AT2G31040	8.98	7.08	1.897	0E+00	10.34	8.72	1.615	0E+00	8.68	10.18	-1.496	0E+00	7.33	7.92	-0.593	2E-01
CATMA5A50147	AT5G54280	11.05	9.29	1.756	0E+00	12.14	10.54	1.595	0E+00	12.74	11.03	1.711	0E+00	9.51	9.47	0.044	1E+00
CATMA3A47440	AT3G54500	9.16	7.44	1.721	0E+00	11.51	9.05	2.460	0E+00	11.02	11.43	-0.408	1E+00	9.22	10.76	-1.542	0E+00
CATMA5A15570	AT5G17300	11.77	10.09	1.683	0E+00	11.84	11.51	0.331	1E+00	8.97	12.34	-3.371	0E+00	7.56	8.47	-0.909	0E+00

Figure 4.3: A table displaying for each dye-swap of the experiment and from left to right, the log2-intensities for sample 1 and 2, the log2 ratio and the Bonferroni P-value. In the example, results are sorted by the ratio values in the dye-swap between the leaf sample T12, extracted at 12 hours after the start of the experiment, and the leaf sample T9, extracted at 9h.

When CATdb was published in 2007, it contained 46 projects with 1 380 hybridizations corresponding to 523 different samples. To date, CATdb contains 216 projects with 8816 hybridized samples in public access out of a total of 338 projects with 14376 hybridized samples. Moreover CATdb includes 39 projects on transcriptome from other species acquired with Affymetrix microarrays (total 496 hybridized samples). Actually, we prepare an adaptation of CATdb to include new types of data (tilling array, RNAseq data, ...) to accompany the platform which takes in charge an increasing number of technologies and species. A modification of a database always takes time so it is important to measure its gain with respect to the future use of these data. It requires a lot discussions between the platform, the bioinformaticians and I because the use of these data depends on the point of view of each of us.

We also collaborate with the team of P. Zimmerman (ETH-Zrick) to transfer transcriptome from CATMA experiments into Genevestigator (Grennan, 2006). We provided more than 1600

normalized intensities that will be available in Genevestigator after their curated check. The statistical part of this collaboration was to provide the normalized intensities of the CATMA experiments in a manner allowing a direct comparison with the Affymetrix data already stored in Genevestigator.

4.1.3 Statistics for the transcriptomic platform

The first manager of the transcriptomic platform of URGV, Jean-Pierre Renou, was hired the same year as I. We arrived in February 2003 and we managed to hybridize and analyze the first data in August. During these first seven months, I learned the basis of the molecular biology, discovered the microarray technology and its specificity and, finally devised a procedure to normalize two-color microarray data and to perform a differential analysis between two modalities of one factor when technical and biological replicates are available.

All this work was done in close collaboration with members of URGV and members of the team "Statistics and Genome" of UMR MIA 518 to be able to propose an efficient analysis of the CATMA data. To be able to propose solutions in the data analysis without doing only this activity, it is important to bring out the genericity of the biological questions because it is absolutely not possible to perform a per-project analysis. In the following, I describe the different choices made to propose statistical pipelines for the transcriptomic platform. To propose a document for the biologists, we wrote an introduction to the statistical methods for microarray data analysis in Martin-Magniette and Robin (2004).

Discussion for the experimental design With Jean-Pierre Renou, we met the collaborators at the beginning of each project to precise the biological question, the expected results in order to propose the most adapted experimental design. It was really important to invite collaborators before the beginning of the genomic project to explain the specificities of the high-throughput technologies and to take them into account in the biological questions. For the majority of the projects, one factor with two modalities is studied. The remaining question is thus the number of biological replicates and the answer is usually guided by the cost of the experiment. If the factor has more modalities, it usually suffices to discuss with the collaborators to organize the comparisons.

After the first year, Jean-Pierre was able to discuss alone the experimental design with the collaborators and I intervened only to validate the experimental design or to construct it in case of ambitious projects.

With the arrival of the high-throughput sequencing technology, various biological questions can be addressed and imply adapted tools and methods in bioinformatics and statistics. For that reason, we invite again the collaborators as soon as the project begins to discuss all the aspects.

Even if these discussions take a long time, they are important and also interesting. It is a very pleasant manner to learn biology. I profit also these discussions to identify some recurrent questions that may require methodological developments.

Data normalization for two-color microarray As I described in Section 3.2.1, there exists technical biases that can be controlled by a rigorous protocol and the others which require a statistical correction. The first step for the CATMA normalization was to identify the technical biases, to classify them in the two bias categories and to propose solutions to correct them.

The first step was to convince all the members of the platform that a rigorous framework is the best solution to remove a majority of the technical biases. After numerous discussions, they agreed to define a fixed protocol and to apply it without any modifications: each project is done by only one person, the quantities of labeled samples deposited on the microarray and the scanner adjustment are fixed within a project and also between the projects. Over time, we believe that these decisions are important to generate comparable data. At first glance, it seems easy to have a fixed protocol for all the members of the platform but this decision can be also felt

as a questioning of their expertise. In 2003, such rules were rarely established on transcriptomic platforms and it was thus important to discuss in order to argue that these decisions improve significantly data quality.

For the uncontrolled biases, the main difficulty is their identification because some of them are specific to the platform. After searching through all biases that might exist and tested several normalization methods, I proposed to use the logarithm of median feature pixel intensity at 635 nm (red) and 532 nm (green) wavelengths without a background subtraction. Indeed the background is calculated from intensity of pixels in the neighborhood of each probe of the microarray and its subtraction can have an impact of the expression difference as shown in Figure 4.4. Then a global intensity-dependent normalization is performed using the lowess procedure (Yang *et al.*, 2002) to correct the dye bias. Because on the first versions of CATMA microarray, probes were organized in three blocks, I proposed also to correct a block effect corresponding to print-tip, washing and/or drying effects. As shown in Figure 4.4, it can be sizeable[2]. The correction was performed by subtracting the log-ratio median over the values for the entire block from each individual log-ratio value.

At the end of the normalization procedure, a normalized log-ratio, i.e. an expression difference (in log base 2) between the two samples co-hybridized on the same array, is given for each probe. A normalized logarithm intensity for each sample is also calculated. This is done according to the within-array correction proposed by Yang and Thorne (2003).

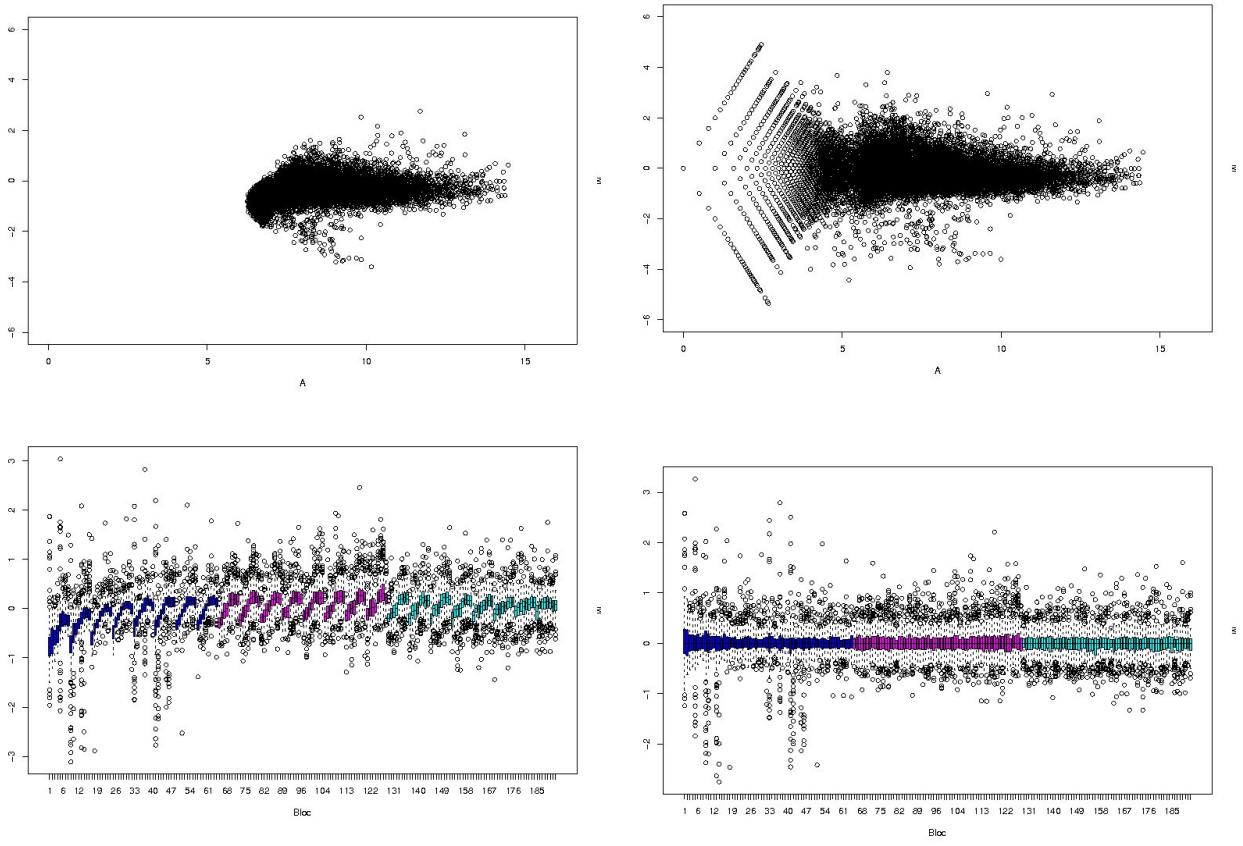

Figure 4.4: **Top left:** MA plot applied on the raw data: the expression difference on the y-axis is described as a function of the averaged intensity on the x-axis. **Top right:** MA plot applied on the raw data with a background subtraction. **Bottom left:** Boxplot of the log-ratio per block after the lowess normalization **Bottom right:** Boxplot of the log-ratio per block after the block correction.

[2]The microarray used to produce the plot was between the first microarrays generated on the platform and these washing and drying problems had been corrected hereafter.

Data normalization for Affymetrix experiment Affymetrix microarray is a one-color microarray where a set of about 10 probes is designed in the 3' UTR of each gene. The normalization of such data is thus different. It consists first in removing the technical biases, which are mainly a non specific signal and an optical noise. Second a between-microarray normalization is made so that intensity signals are comparable. Third the hybridization signals of the probe set are summarized into an hybridization signal for the associated gene. When the platform decided to use Affymetrix technology, many works were already published for the normalization of these data. Consequently, Julie Aubert (UMR AgroParisTech/INRA MIA 518) and I reviewed the different methods and finally decided to use the RMA or GC-RMA method (Irizarry *et al.*, 2003; Wu *et al.*, 2004) according to the presence of not of mismatch probes on the microarray.

The main drawback of the one-color microarray is that it is impossible to separate the sample effect from the slide effect and due to this effect mix-up, it is not possible to use the mean squares to evaluate the normalization performance. It is also difficult to evaluate the impact of the between-microarray normalization, usually based on the quantile method (Bolstad *et al.*, 2003). For example, starting from the same files of raw data of two conditions measured on two biological replicates, the number of genes differentially expressed could vary between 0 and 500 if the quantile method is applied by condition or on the overall samples. For all these reasons, I do not like working with one-color microarrays and I convinced the platform to choose two-color microarray as long as possible.

Choice for the differential analysis The goal of the differential analysis is to determine genes for which expression differs between conditions. The differential analysis is done using a test but the number of observations per gene being small, it is inadequate to calculate the test statistic from the empirical variance of the expression difference for each gene.

In 2003, when I had to devise a method of differential analysis, few methods existed to take into account the small number of replicates. I tested various methods and based on results obtained on experiments done at URGV, we decided to use a paired t-test where a common variance of expression difference is calculated over the probes not displaying an extreme specific variance (too small or too large). The excluded probes are defined as those with a specific variance/common variance ratio smaller than the α-quantile of a chi-squared distribution of one degree of liberty or greater than the $1-\alpha$-quantile of a chi-squared distribution of one degree of liberty, with $\alpha = 0.0001$. This rule stems from a direct application of Cochran's theorem.

When technical and biological replicates are available, I proposed to perform the differential analysis on the average over the technical replicates. This solution takes into account both variabilities by assuming that the highest is the biological variability.

Microarray data is an example of multiple test framework. Raw P-values have to be adjusted to control the false-positives. With the proposed variance modeling, I decided to control the Family Wise Error Rate (FWER) by using the Bonferroni method.

With the new version of CATMA where probes are synthetized oligonucleotides, I changed the method of differential analysis. We opted for Limma with a FDR control.

Choice for the clustering After the differential analysis, the next step is usually the clustering of the log-ratios over several experiments to determine co-expressed genes. I think that the use of the clustering methods depends strongly on the biological question, so I advice the software Genesis to our collaborators so that they performed their clustering.

Development of pipelines For the implementation I had originally planned to use functions proposed by the Bioconductor project, but I quickly abandoned this idea. The functions in Bioconductor are designed for standard analyzes and modifications are not easy. Moreover I wanted to take the CATMA specificities into account so I wrote R functions for the normalization and the differential analysis. The idea is to write a script where the members of the platform have to add only some important information. They then launch the R script by a shell command and

analysis outputs are written in files. Consequently, the platform uses the software R indirectly. There exists a pipeline for each type and version of a technology. I teach the developed methods to all the members of the platform who are then able to analyze data themselves. They contact me only if they detect a problem.

4.1.4 Publications associated with the platform activity

The platform pays also great attention that the data generated are properly exploited. Figure 4.5 shows the evolution of the transcriptome projects managed by the platform and evolution of the associated publications. The platform has a total of 94 publications between 2004 and 2012 and I was included in the list of the co-authors 15 times for different involvement as explained below:

Lurin *et al.* (2004) is the first paper using transcriptome data produced with a CATMA microarray. It was the first description of the statistical analysis of CATMA data. Additionally I also performed some tests to better characterize some subsets of genes with respect to the whole set of Arabidopsis genes.

In Jammes *et al.* (2005), Herbette *et al.* (2006) and Ramel *et al.* (2007) are among the first papers using CATMA microarrays. Ruffel *et al.* (2008), Cossegal *et al.* (2008), Elis *et al.* (2009), Elis *et al.* (2008) and Dubois *et al.* (2011) are among the first papers using Affymetrix microarray on the transcriptomic platform of URGV. For these eight projects, I was involved in the experimental design, the statistical analysis, the result interpretation and the manuscript writing.

To date, the biologists are more familiar with the transcriptomic data and their interpretation. Moreover, the platform provides a detailed description of the statistical analysis that can be directly included in the Material and Methods section of a manuscript. Consequently my statistical developments become tools and I am less involved in the projects, except when the reviewers have statistical comments. It is the case for Reymond *et al.* (2012) and Moreau *et al.* (2012).

The platform was also involved in large transcriptome projects with more than fifty microarrays. Cohen *et al.* (2010) deals with a comparative transcriptomics of drought responses in two genotypes of Populus. I was in charge of the experimental design, in particular for the organization of the trees in the greenhouse and the definition of a process for implying the drought stress. I participated also to the meta-analysis of the data. Rengel *et al.* (2012) is based on a large transcriptomic projects studying eight sunflower (Helianthus annuus) genotypes to two drought stress scenarios. I was mainly involved in the normalization of the Affymetrix data (normalization of a subset of probes available on the Affymetrix microarray and modification of the normalization to include a new technical effect impacting the differential analysis). Cubillos *et al.* (2012) is the first project of eQTL based on two-colors microarrays in dye-swap. A total of 314 recombinant inbred lines (RILs) were studied. I was in charge of the experimental design and used a random pair design for each comparison of two recombinant inbred lines (RILs). I performed initial data analyzes to provide a normalized hybridization signal for each of the 314 RILs. Recently I initiated a collaboration where data produced for Cubillos *et al.* (2012) were used to infer a regulatory network using Bayesian networks (Vandel *et al.*, 2012).

My expertise on the transcriptome data analysis being completely independent of the organism, I also worked on transcriptome projects external to the URGV platform. In these collaborations, I did not perform the analysis myself, just helped to choose the most adapted model (Abdelkarim *et al.*, 2011; Chadi *et al.*, 2010; Teixeira *et al.*, 2009; Navajas *et al.*, 2008).

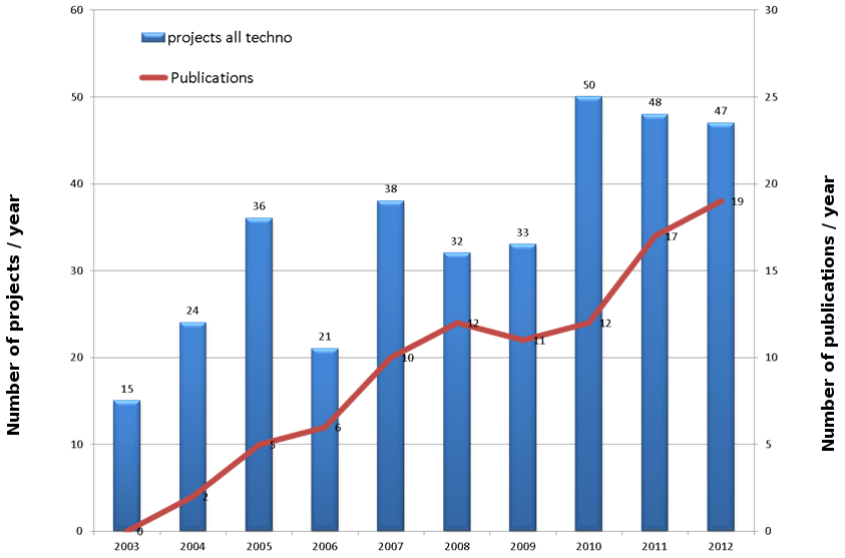

Figure 4.5: Evolution of the transcriptome projects by the platform and evolution of the associated publications.

4.2 Use of the transcriptomic resource to improve the structural annotation

4.2.1 Background

The model plant *Arabidopsis thaliana* was sequenced and annotated in 2000 and a new annotation release is proposed each year taking into account the completion of the genome sequence and the improvement of gene predictions. In addition to this global semi-automatic annotation, different works also improved Arabidopsis gene detection using orphan ESTs, comparative genomics, or combination of data through expertise of gene families. In Aubourg *et al.* (2007), we contributed to the structural annotation improvement by analyzing the genomic resource generated by the transcriptomic platform of URGV.

Based on the official TAIR version 6.0 annotation release, among the 24576 probes of CATMA microarray, 677 probes mapped outside genes. We selected them as novel candidate genes and investigated their expression to reveal possible under-predicted functional genes in Arabidopsis.

4.2.2 Characterization of novel genes

Transcriptome of 522 different samples covered numerous developmental stages, biotic and abiotic stresses and mutants were extracted from CATdb. The intensity signal of a sample was defined as the average of the normalized logarithm intensities of the two arrays of the dye-swap. For each sample, I defined an hybridization threshold using MixThres, described in Section 3.3.1, with $\varepsilon = 10^{-4}$. Among the novel candidate genes, 465 showed hybridization in at least one sample. To validate the results, two different experimental approaches were used. First a RT-PCR was performed using 4 different mRNA samples from roots, leaves, flowers and pollen for each of the 465 genes. Second, we sequenced amplicons obtained for 411 among the 465 genes. For 410 putative novel genes among the 465, we obtain a proof of transcription by both experimental approaches.

At a structural and functional level To further characterize the newly identified genes, additional data-mining was performed. All these structural and functional information is sum-

marized on figure 4.6 and in the following:

- Other independent evidence of transcription was found for 204 genes.
- For 215 genes, significant sequence similarities at the protein level were detected at least in one other locus in the Arabidopsis genome and/or with proteins from different species, indicating that they belong to known gene families.
- Inference of function by similarity could be made for only 71 genes and the remaining 394 genes encode proteins with unknown biochemical function.
- Surprisingly, 86 genes previously annotated at the BAC scale were ignored in the whole genome annotation done later, probably because of poor supporting data.
- The topological distribution of the 465 novel genes was quite similar to all the Arabidopsis coding genes. They were evenly distributed in the 5 chromosomes and were rarely present in the peri-centromeric regions or other identified heterochromatic regions.

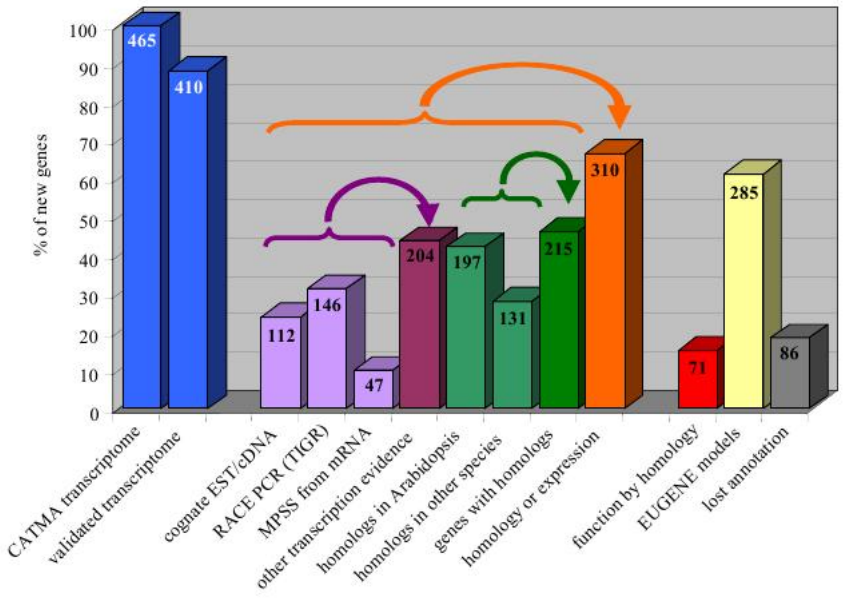

Figure 4.6: The validated transcriptome fraction is the result of our RT-PCR approach. The other evidence of transcription comes from cognate EST/cDNA, RACE PCR from TIGR or MPSS data (purple columns). They are summed up in the other transcription evidence class. The fractions of the novel genes sharing similarities with other genes (in Arabidopsis and/or in other species) are indicated in green. The orange column highlights the fraction of novel genes for which there is an indication complementary to CATMA data (homology or transcription) of the gene presence.

The newly discovered genes were mainly characterized by their short size with an average of 411 bp compared to 1247 bp for the already known Arabidopsis genes (Figure 4.7). This result could explain why these genes were missed by automatic annotation. Indeed, their coding potential (unusual length surrounded by larger intergenic regions) may be difficult to detect by a semi-HMM. Beyond this, our approach also detected large conserved genes with 9 and 11 exons.

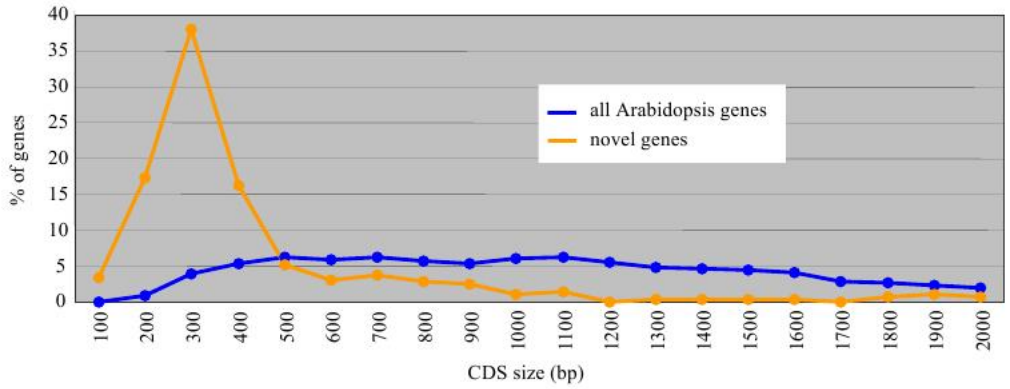

Figure 4.7: Relative distribution of CDS size (in bp) of all Arabidopsis genes (in blue) and of the novel genes (in orange) for which Eugene v1.59 has predicted an intron-exon structure.

Expression of the novel genes After the structural characterization, we investigated the expression of these novel genes. It was done through a comparison of the transcription data obtained from 522 hybridized samples for the 465 novel genes and all the 21 260 Arabidopsis genes having a CATMA probe. The results are the following

- Most newly identified genes are detected in a limited number of experimental conditions. Indeed, 40% of the novel genes are detected in 1 to 5 mRNA samples while there are only 16% of all the Arabidopsis genes in this category.
- Only 5% out of the 465 novel genes are detected in more than 150 samples. This number is very low compared with the 28% of all the Arabidopsis genes that are detected in the same number of samples.
- The novel genes were never detected in more than 95% of the 522 samples. Thus, they do not belong to the category of constitutively expressed genes also frequently referred to as housekeeping genes.
- We found 103 novel genes for which expression is reported in only one organ. Even if we cannot conclude that there is complete organ-specificity from our data, the transcription of these 103 genes is clearly highly preferential in only one organ.

Explanations about the rarely observed transcription of several novel genes Three explanations that are not fully exclusive may be given to the rarely observed transcription of several novel genes: (i) we may consider that some probes give an artefactual signal in one hybridization. Nevertheless, we confirmed the transcription of 88% of the novel genes by RT-PCR and sequencing. (ii) it is possible that the transcriptome approach allows the detection of an expression of genes constitutively expressed at low level. Indeed, constitutive genes always expressed at low level would generally give hybridization signals below the thresholds for considering the corresponding probes as hybridized. It is only in a small number of samples when the expression is just slightly higher than in all other samples that the probes corresponding to these genes would be recognized as hybridized by MixThres. (iii) the signal might depend on relatively rare physiological or environmental situations.

We tried to evaluate the relative explanatory potential of the last two expectations by comparing the distributions of intensity signals for both the whole genome and the novel genes. We expected that genes expressed constitutively but at low level would present a maximum hybridization signal lower than the genomic distribution. However we observed that the novel genes show the same relationship between the number of samples and the maximum hybridization signal as the whole genome does (Figure 4.8). There was no novel gene for which the

maximum intensity signal distinctly departs from the known genes showing the same number of hybridized samples. Thus, all together the transcriptome data for the novel genes suggest that the transcription of several of these genes were not only organ specific but also more specific to rare endogenous or environmental conditions than the whole genome. This double control of transcription might well explain our observation of transcription of several novel genes in only one biological sample. For this reason, the transcripts corresponding to these genes are less often present in the cDNA libraries which, in Arabidopsis, cover several organs but relatively few different environmental conditions.

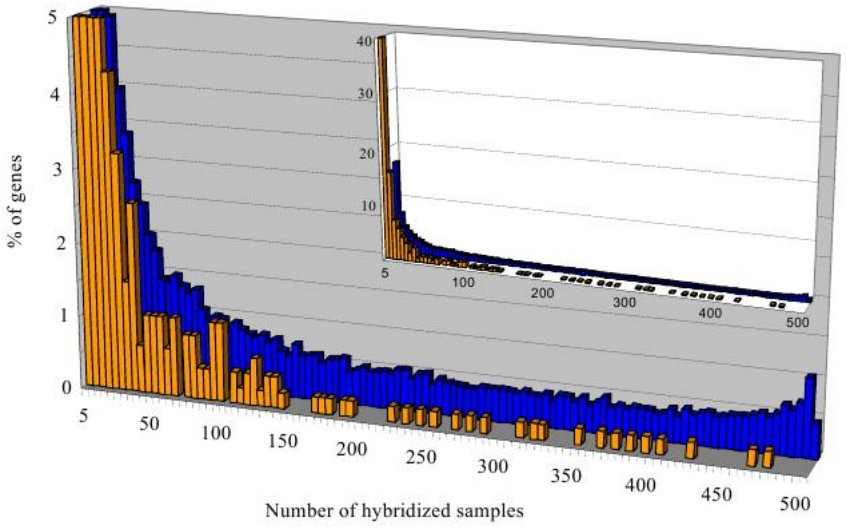

Figure 4.8: Each gene is spotted according to the number of samples in which it has been detected and to the hybridization signal (transcriptional activity): The minimum intensity in any of the hybridizations is in blue and orange for all the Arabidopsis genes and the 465 novel genes respectively. The maximum intensity is in purple and yellow for all the Arabidopsis genes and the novel genes respectively.

Detection of erroneous gene merging We also used our analysis to verify annotation of 422 loci. According to the TAIR annotation, distinct GSTs on these 422 loci matched a single gene (not supported by full-length cDNA) whereas the software Eugene predicted two different gene models (Schiex *et al.*, 2001). For 13 loci, the transcriptome results showed that two GSTs associated with the same gene provided opposite ratios in the same transcriptome comparison. We thus concluded that they matched two different genes. An example is reported on Figure 4.9.

4.2.3 Conclusions

This work was a noticeable step forward in the improvement of the Arabidopsis genome annotation. We increased the number of Arabidopsis validated genes by 465 novel transcribed genes to which we associated several functional annotations such as expression profiles, sequence conservation in plants, cognate transcripts and protein motifs. All these results strongly illustrate that the annotation process is a long and difficult task and that many years are necessary after the first release of the sequence of a complex eukaryote genome to obtain (nearly) full knowledge of its gene content.

Personally, I learned with this project about the manner to interpret the statistical results of a genomic project. I understand that once a gene subset of interest is defined, several measures can be calculated on it and these measures have sense only if they are compared to the same measures

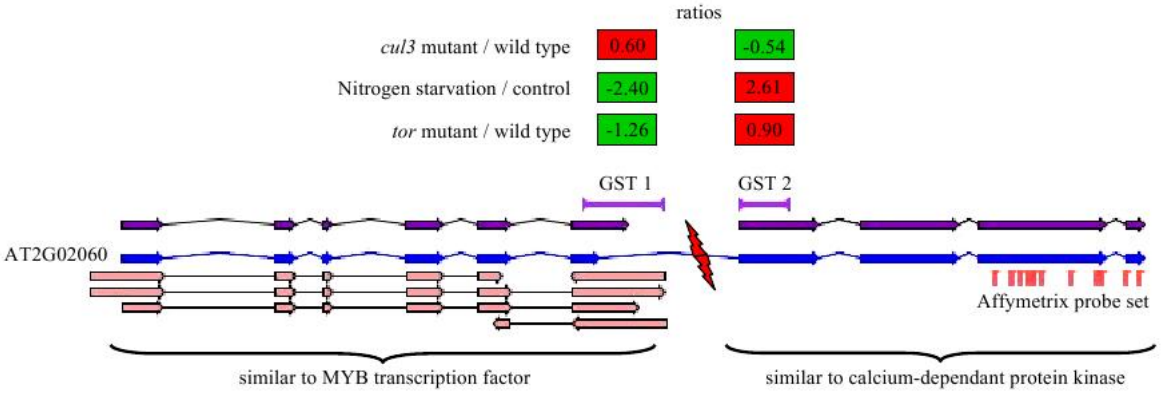

Figure 4.9: Erroneous gene merging occurred in the annotation process and detected using CATMA transcriptome data. The gene AT2G02060 is the fusion of two genes encoding a MYB transcription factor and a calcium-dependent protein kinase. In three independent experiments the two GSTs provide significant ratios (Bonferroni, P-value ≤ 0.05) indicating that the first gene is up-regulated while the second one is down-regulated in the same comparison. Cognate EST and cDNA supporting the MYB gene are represented by pink arrows.

calculated on the whole gene set. Only then a measure become a characteristic. Consequently in such projects there are two distinct statistical analyzes and both require different skills. The project is the first one where I was involved in both analyzes. In the other projects described below, I extracted only information from the experimental data.

4.3 Other projects

Additionally to my activity on the transcriptomic platform of URGV, I also collaborated with other teams of URGV for projects not dealing with omic data. In this case, the aim is to apply correctly classical methods of statistics. The second type of described projects gathers together the applications of MixThres and MultiChIPmix on chIP-chip data.

4.3.1 Projects not dealing with omic data

A recurrent question in biology is to determine if an observed phenomenon in a given experiment can be observed by chance. In this case, my work consists in writing the biological question into an hypothesis test. In Faivre Rampant *et al.* (2011), I applied such strategy on a project about the composition of the oak genome. Starting from the number of match of a set of sequences against the oak genome, the aim was to determine the number of match required to consider that a sequence is a repeat elements.

In a second project, I helped Cléa Houel, a PhD student at URGV, to perform the statistical analysis to characterize the genetic variability of berry size in the grapevine (Houel *et al.*, 2013). We used classical tools of statistics (tests and hierarchical clustering).

4.3.2 Projects on control mechanisms of gene expression

Background Stable repression of gene expression is an important aspect of the developmental programs of higher organisms. In plants and animals, DNA is organized within chromatin. Schematically, chromatin is found in two varieties: euchromatin and heterochromatin. The euchromatin is a lightly packed form of chromatin that is rich in gene concentration, and is often under active transcription. The heterochromatin is a tighter packing where genes are repressed. Besides the chromatin contains at its core a set of proteins called histones. These

proteins can be modified for example by methylation or acetylation of lysines and their states influence the chromatin variety and consequently have an impact of the gene expression. With ChIP-chip experiments, it is now possible to investigate the interaction protein-DNA at the genome scale and I collaborated with two different teams involved in the control mechanisms of gene expression of *Arabidopsis thaliana*.

To identify enriched regions corresponding to targets of the protein under study, MixThres, described in Section 3.3.1, was used in the two first projects (Benhamed *et al.*, 2008; Turck *et al.*, 2007) but the application of MixThres differs in both projects because the modeling was not ill-suited. Nevertheless these analyzes allow us to really understand the data specificities and we developed MultiChIPmixHMM, described in Section 3.3.2. Through a collaboration with the team of Vincent Colot, I applied MultiChIPmix on various chromatin marks and was involved in two important projects dealing with the control mechanisms of gene expression. Only a part of the biological results of the four analyzes are described in the following.

Targets of the histone acetyltransferase GCN5 Benhamed *et al.* (2008) focused on a study of the protein GCN5 known to play an essential role in chromatin modication and transcriptional regulation. In order to gain deeper insight into the nature and chromosome-wide distribution of GCN5 in the Arabidopsis genome, they performed chIP-chip experiments with a microarray where probes cover approximately 20 000 *Arabidopsis thaliana* promoter regions. GCN5 was studied in a wild-type and in two mutants where the gene *gcn*5 is modified to investigate the importance of a domain, named GCN5 bromodomain. Besides they also studied acetylated H3K14, known as a target of GCN5 in the two mutants to determine if the GCN5 bromodomain was important for acetylation of H3K14. For each dataset, I determined an hybridization threshold for the IP values using MixThres with $\varepsilon = 1e - 4$.

ChIPchip results showed that, in wild-type plants, GCN5 binds to a large proportion of the promoters (3352/8471, i.e. 40%). Target promoters were distributed over all ve chromosomes. Chromatin isolated from gcn5-1 and gcn5-2 plants had comparable proles. Comparative analysis revealed that 89% of the GCN5 target promoters bound in wild-type plants were precipitated from the chromatin of both mutants, indicating that the truncated GCN5 proteins were still recruited to a large proportion of the target promoters This result implies that the remaining promoters that were targeted in the wild-type, but not in the mutants, require the GCN5 bromodomain for binding. Result interpretation also suggests that the GCN5 bromodomain is necessary to acetylate H3K14 in a subset of promoter regions.

To determine whether GCN5 binding affected transcriptional regulation, we characterized the transcriptome of wild-type, gcn5-1 and gcn5-2 plants using CATMA microarrays. Based on the differential analysis results, about 3.5% of the genes were differentially expressed in both mutants when compared to the wild-type, a percentage close to the 4.0% already found in a previous study.

Protein TFL2/LHP1 Turck *et al.* (2007) focused on a study of the protein TFL2/LHP1 known to play a role to maintain the stable repression of genes. In order to gain deeper insight into the nature and chromosome-wide distribution of TFL2/LHP1 target sites and their associated histone marks, they performed chIP-chip experiments with a DNA tiling array of the entire Arabidopsis Chromosome 4. Four chromatin marks were also studied.

To identify the targets of the proteins, I devised a procedure with Vincent Colot. I applied first MixThres on IP intensities to select only probes for which IP value is higher than the hybridization threshold. MixThres is then applied on the logratio (IP/Input) of these probe subset to identify the target of the protein. Figure 4.10 is an example for H3K27me3.

All the enriched regions were then interpreted from a biological point of view and visualized on a genome browser (Figure 4.11). We demonstrate that TFL2/LHP1 associates with hundreds of targets across this chromosome, the vast majority of which correspond to genes located within euchromatin. Further- more, we show that TFL2/LHP1 associates almost exclusively and

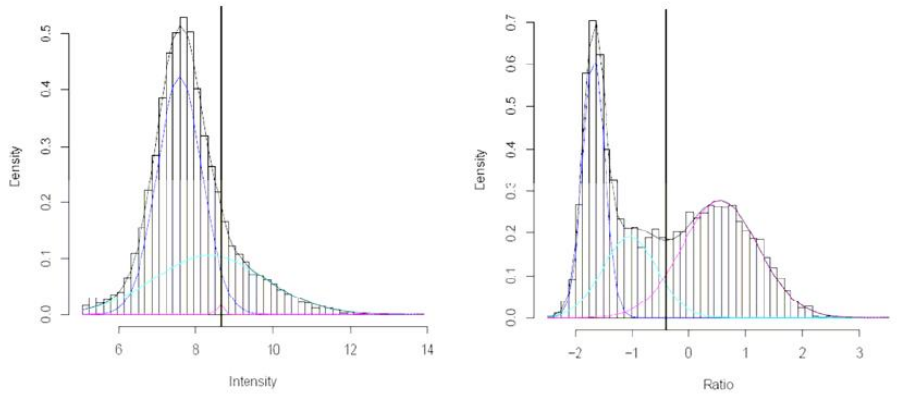

Figure 4.10: Histogram of IP values for the second biological replicate of the mark H3K27me3

nearly co-extensively with H3K27me3. Moreover, the absence of noticeable changes in the distribution of H3K27me3 along Chromosome 4 in lhp1 mutant plants indicates that TFL2/LHP1 is not involved in the deposition of this mark. Rather, TFL2/LHP1 specically associates with H3K27me3 in an in vivo context, indicating that it is involved in a general mechanism of gene regulation mediated by Polycomb repressive Complex 2.

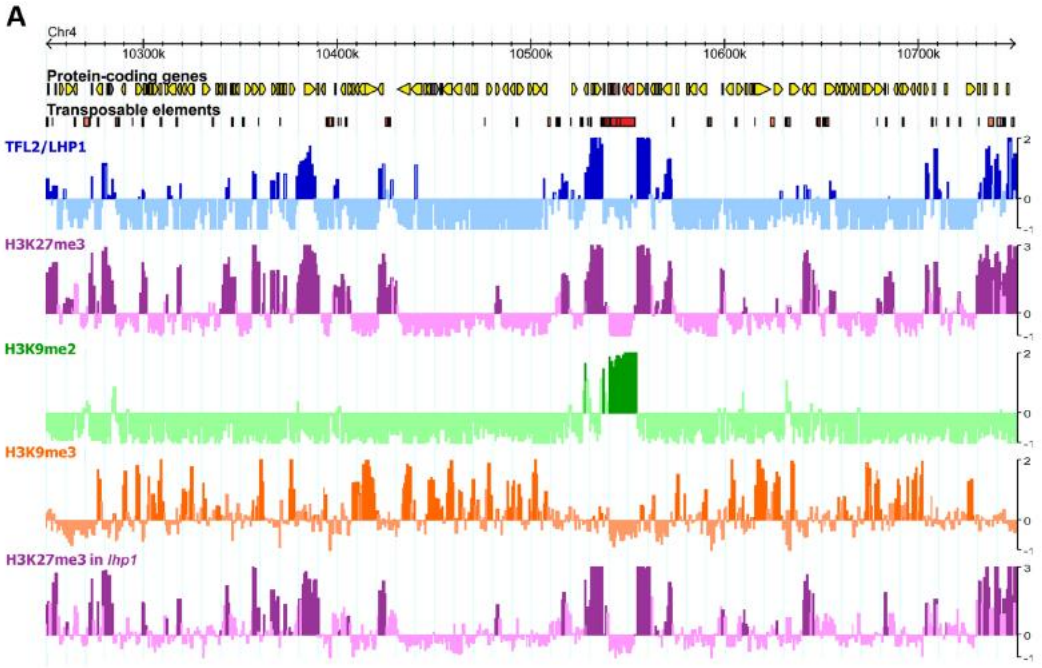

Figure 4.11: Genome browser view of a 500-kb region of Chromosome 4 (positions 10.25 Mb to 10.75 Mb) showing the co-extensive association of TFL2/LHP1 with H3K27me3, and the lack of overlap between H3K9me2, H3K9me3, and H3K27me3. IP/INPUT ratios (log2) reporting significant association are marked in dark colors.

Epigenomic map Roudier *et al.* (2011) generated epigenomic maps for eight histone modications (H3K4me2 and 3, H3K27me1 and 2, H3K36me3, H3K56ac, H4K20me1 and H2Bub) in the model plant Arabidopsis to explore the principles governing the combinatorial association of

chromatin marks along the genome. I was in charge of the analysis and determined the enriched region for each mark using MultiChIPmix.

Based on these enriched regions detected with MultiChIPmix and already known regions for H3K9me2, H3K9me3, H3K27me3 and DNA methylation, they provided an epigenomic map of *Arabidopsis thaliana*. These 12 chromatin marks, which collectively cover 90% of the genome, are present at any given position in a very limited number of combinations. Moreover, they showed that the distribution of the 12 marks along the genomic sequence defines four main chromatin states, which preferentially index active genes, repressed genes, silent repeat elements and intergenic regions.

Additive inheritance of histone modications in Arabidopsis thaliana intra-specic hybrids Little is known about the stability of these epigenomes when distinct genomes are brought together by intra-species hybridization. In Moghaddam *et al.* (2011), the genome-wide distribution of histone modications H3K4me2 and H3K27me3 in the inbred parental accessions Col-0, C24 and Cvi and their hybrid offspring was compared by chromatin immunoprecipitation in combination with genome tiling array hybridization. ChIP-chip analysis was done with MultiChIPmix and results for the parental accessions are summarized in Figure 4.12.

This analysis revealed that chromatin modication variations exist among accessions of A. thaliana. However H3K4me2 and H3K27me3 were rather stable in response to intra-species hybridization, with mainly additive inheritance in hybrid offspring. In conclusion, intra-species hybridization does not result in gross changes to chromatin modications.

Table 2 H3K4me2- and H3K27me3-associated domains in Col-0, Cvi and C24

					Domain size (kb)			Annotation	
Experiment[a]	Genotype	Modification	Associated tiles (from total of 717 235)	Number of domains	Maximum	Minimum	Mean	Genes	TEs
A	Col-0	H3K4me2	287 285	21 539	31.2	0.3	2.2	22 961	2063
	Cvi	H3K4me2	281 000	20 972	21.9	0.3	2.2	22 489	2005
B	Col-0	H3K27me3	140 035	11 182	22.7	0.3	2.1	9125	4950
	Cvi	H3K27me3	138 791	12 585	20	0.3	1.9	9326	5357
C	Col-0	H3K27me3	147 318	9834	26.7	0.3	2.5	9427	4704
	C24	H3K27me3	125 947	10 068	22.5	0.3	2.1	8118	4471

[a]Chromatin preparations for experiments A and B were performed in parallel but the one for experiment C was performed independently.

Figure 4.12: H3K4me2- and H3K27me3-associated domains in Col-0, Cvi and C24

Chapter 5

Future projects on genomic networks

Since March of this year, I am the new leader of the team "Bioinformatics for the predictive genomics" at URGV. In this new context, my first task was to write a team project in the framework of the creation of a plant institute on Paris-Saclay campus. This is the project that I describe below where research projects in statistics are proposed in a genomic framework in order to contribute to the functional and relational annotation improvement of *Arabidopsis thaliana*. The bioinformatic part of the team project is not detailled.

5.1 Background

Despite the increasing number of fully sequenced genomes and the technical progress in experimental genomic, the functional annotation process is still powerless for 15 to 40% of the predicted genes (Gollery *et al.*, 2006). Up to now, this limit is true from bacterial to mammalian genomes. Among the whole plant genomes published, the *Arabidopsis* genome annotation has the advantage to benefit, since its first release 13 years ago, from important efforts including automatic, experimental and curated approaches. Nevertheless, after 10 official versions of annotation, around 4000 genes are still reported as encoding unknown proteins by the absence of homologs with known function in any other plant or organism (Lamesch *et al.*, 2012). Furthermore, only 14% of genes have a biological function characterized by experimental approach. All together, it shows that the gap between the structural and functional annotation remains wide.

With the availability of thousands of transcriptomes, it is now possible to shift from a gene by gene approach to a faster and more global genome investigation. As shown in Section 4.2, a global exploitation transcriptome resources contributes to the structural annotation improvement. Additionally, transcriptome resources also contribute to functional annotation improvement (Eisen *et al.*, 1998). The underlying hypothesis is that a cluster of genes sharing similar transcription profiles is enriched in genes involved in a specific biological process, which can be tentatively assigned to orphan genes present in the cluster. Such an approach has been applied to Arabidopsis and reported for punctual gene function characterization Persson *et al.* (2005) and more general functional classification of orphan genes Horan *et al.* (2008).

Our team is devising a such approach by targeting the genes involved in the response to various stresses. Our approach is based on 445 transcriptome comparisons, extracted from CATdb. Among these comparisons organized in 17 types of stress, more than 18000 genes are identified as differentially expressed at least once. To perform a co-regulation network, I applied SRUW model, described in Section 2.1.2, on each type of stress. The number of considered genes per stress varies between 1500 and more than 14 000 and the number of co-expression clusters varies between 13 and 60 among the 17 types of stress. The first results of the co-expression analysis provides many very exciting insights and a rigorous biological interpretation is in progress. This is the subject of my new PhD student, Rim Zaag, who began one year ago.

Despite these hopeful results, many biological aspects have been neglected due to technical difficulties. It concerns in particular the co-expression analysis of non coding protein genes

for which few transcriptome measurements are available. It concerns also the integration of RNA-seq data. These two questions raise methodological problems that I present in Section 5.2.

A second complementary question is to progress on the definition of regulatory networks. To date, the most used approaches are genetic approaches where genes are mutated one-by-one to identify their partners. Hence it takes a long time and the development of methodological approaches to speed-up the inference of regulatory networks are welcome. During the past ten years, many statistical models describing the relationship between genes were developed, as well as dynamic Bayesian networks, Gaussian graphical models, or networks based on the concept of mutual information. Gaussian graphical models are the most frequently used models when only transcriptome data are available (Chiquet *et al.*, 2009; Giraud *et al.*, 2012). These models assume a Gaussian distribution of gene expression levels observed in several conditions and describe the dependence structure between genes using the inverse of the covariance matrix, called the precision matrix. Null entries of this matrix correspond to pairs of genes conditionally independent. Conversely, the non-zero entries characterize the direct relationships between genes. Under the assumptions that the list of genes is known and that the level of expression is constant across all conditions, these models have good statistical properties when the number of experiments and the number of genes are reasonable (Verzelen, 2012). Nevertheless their application to transcriptome data is not straightforward and raise methodological questions that I present in Section 5.3.

5.2 Open questions for improving the co-regulation networks

The associated methodological problems can be viewed as questions of modeling, model selection and, algorithms. All the projects dealing with RNA-seq analysis will be done in the project MixStatSeq coordinated by C. Maugis-Rabusseau between 2014-2018.

5.2.1 Modeling challenges

Observations with missing values not at random Because the probes tagging the predicted small RNA genes and the non coding protein genes were spotted on a recent version of CATMA array, the largest number of hybridizations performed on the previous versions does not give information on their transcription. Consequently they could not be considered as observations with values missing at random and they could not be included in the co-expression analysis. However, they play an important role at the genomic level and they should be included in the analysis of co-expression.

A first idea is a pragmatical solution. It consists of assigning the predicted small RNA genes and non coding protein genes in the identified co-expression clusters by using the experiments where coding-protein genes, predicted small RNA genes and non coding protein genes are measured together. However I presume that in many cases, the conditional probabilities will not be sufficiently informative to assign these genes in a single cluster.

A second pragmatical idea is to study the predicted small RNA genes and non coding protein genes independently from the coding-protein genes and to compare their expression profiles to link the two categories of genes.

If both ideas provide insufficient results to study correctly the predicted small RNA genes and the non coding protein genes, then I will work on modeling taking into account observations not missing at random.

Clustering of RNA-seq data As HTS experiments continue to gain in popularity for studies of the transcriptome, it is important to think about the clustering of RNA-seq data. As for the clustering of microarray data, we would like to use model-based clustering.

Our first strategy consists of fitting a mixture model directly to the untransformed RNA-seq data, leading to a final gene clustering that is interpretable on the original scale of the data.

To construct such a model-based clustering procedure, we are first confronted with modeling of RNA-seq data. In the first proposed models (Si *et al.*, 2011; Rau *et al.*, 2011), variables are assumed to be conditionally independent given the mixture component, and the dependency among replicates of a same condition is accounted for in the decomposition of the mean parameter of the Poisson distributions. Nevertheless, these mixture models are not adequate to account for the characteristics of RNA-seq data. Thus, alternative models are needed, such as those based on negative binomial or Poisson-Tweedie mixtures.

Our second strategy will focus on developing more flexible approaches that can better account for multivariate dependencies among variables. To directly model these dependencies, it could be envisaged to consider mixtures of multivariate Poisson distributions; however, these mixtures are impractical when the number of variables increases. Thus, we will investigate a combination of an appropriate data transformation and multivariate model-based clustering approaches such as multivariate Gaussian mixture models. The primary difficulty lies in the choice of transformation and its impact on the downstream clustering result. We propose to consider a kernel approach, which consists of mapping input data through a transformation into a high-dimension Hilbert space (as for example Wang *et al.*, 2003). The clusters obtained from the transformed data will certainly not be interpreted in the same way as those obtained with the first strategy. Thus, an additional challenge will lie in comparing the gene clusters obtained in each of the proposed strategies, and selecting the most appropriate strategy for a given set of data; for this aspect, exchanges with biologists will be crucial to interpret the clusters in each case.

Clustering of RNA-seq data with variable selection It is important to anticipate the existence of increasingly large RNA-seq experiments in a near future. This information growth implies the crucial problem of variable selection to improve the clustering and its interpretability.

The first strategy consists of considering variable selection without preliminary variable transformation to lead to interpretable results on the original data scale. The main challenge is thus to design an efficient variable selection procedure in adequacy with the biological aims and with the model-based clustering of count data. The idea is to extend the works presented in Section 2.1.2 to the context of gene clustering from RNA-seq data. This consists of developing a model of global variable role (relevant, redundant or noisy variables) in adequacy with the modeling choice defined in the previous paragraph. The main work is to find a function with good properties to link the redundant and relevant variables to keep the SRUW framework.

Clustering of heterogeneous transcriptome data To date, transcriptome data are mainly obtained with microarray data and, in a near future, transcriptome will obtained with HTS technology. This evolution will have an impact on the statistical analysis since the nature of the data is different: with microarray technology, observations are continuous and the measures are on the expression difference between two conditions (logratio), whereas the HTS technology is more sensitive and provides count data representative of a transcript abundance.

The strategy is very pragmatic, it consists in focusing on comparisons of gene clustering results obtained independently from microarray and RNA-seq data collected within the same experiment. It will enable an assessment of the effect of each technology on the downstream results and well as the complementarity between both. This will help us to refine the questions ill-defined for the moment and to structure a future strategy.

5.2.2 Model selection

Once a modeling framework is defined, many parameters have to be fixed. One strategy is to consider a model collection and to develop a penalty function to choose one model or weights to aggregate them. The model selection will be considered in two frameworks:

1. If we develop mixture of Poisson-like distributions, a crucial step will be the choice of the criterion to select the number of clusters and potentially the type of model, if several

are in competition. Two points of view may be considered: an asymptotic penalized criterion, such as the BIC or ICL, or a non-asymptotic penalized criterion, which requires a theoretical study (in particular, to control the bracketing entropy of the considered model collection) and a practical study for calibrating the penalty (see for instance Baudry *et al.*, 2012).

2. We will investigate clustering procedures able to simultaneously account for RNA-seq data and other biological information, including previously obtained microarray data and gene annotation information. The first strategy is to perform the clustering of a single data type (e.g., RNA-seq data) and to construct a model selection criterion accounting for the additional external information. The main challenge is to modify the model selection criterion so that the penalty function reflects the additional types of information. The second strategy is to use a separate clustering procedure for each type of data followed by a method to aggregate information obtained by each.

5.2.3 Algorithms for variable selection

The algorithms developed for variable selection in model-based clustering are greedy and it is necessary to work on their acceleration.

The first idea is to pursue the work of Meynet and Maugis-Rabusseau (2012) to give a definition of a noisy RNA-seq variable for the gene clustering and to propose a l1-type procedure to select a random subcollection of variable subsets. The main difficulty will be the adaptation of such a procedure to the RNA-seq data modeling.

A second idea consists of adopting a dimension reduction point of view, namely to cluster genes in a lower dimensional space. It could be done by using a dimension reduction procedure prior to clustering genes or by providing a model-based procedure which simultaneously accounts for the dimensional reduction and the clustering goal (Bouveyron and Brunet, 2012; Mcnicholas and Murphy, 2008).

5.3 Inference of regulatory networks

Since 2010, I am animating a methodological group (NETBIO) with Matthieu Vignes and Simon De Givry (BIA-Toulouse, INRA) in order to understand the problematic of the regulatory networks from a statistical and biological point of view. The group include 80 members mainly statisticians and informaticians and some biologists. NETBIO is mainly funded by the Department of Applied Mathematics and Informatics of INRA and we organize an annual meeting bringing together about 30 people every year. The two following projects are issued from the discussions of the group.

5.3.1 Statistical learning and multivariate analysis for robust inference

Trung Ha is beginning a PhD this September, co-supervised with J. Chiquet and G. Rigaill.

The most recent comparative studies have clearly demonstrated that the methods of inference cannot infer a network of all the genes of an organism. The inference of a subnet seems to be a more realistic goal. It is therefore crucial to select the right genes. *Ad hoc* methods based on the results of the differential analysis is usually used, but some important genes can be missed and it leads to a poor reproducibility of results.

When comparing different biological conditions, the expression of a gene might shift, either in its average expression level characterized by its mean, or in its interactions with other genes characterized by the covariance matrix. These two types of events are usually analyzed independently even though they are clearly related. In order to alleviate these limitations, we would

like to propose a unified strategy to address these two questions and identify key genes affected either in terms of their mean or their interactions with other genes.

Our strategy has three steps. First considering that genes are independent, the differential analysis question will be rewrite as a problem of penalized likelihood. Second it will consist of relaxing the independence assumption and to consider a known matrix of relationship between genes. Third, the matrix of relationship between genes will be assumed to be partially or completely unknown. The work combines both modeling and algorithmic questions.

5.3.2 High-dimensional regression for studying transcription factors

Yann Vasseur will begin a PhD this September, co-supervised with G. Celeux

In this project, the idea is to focus on the transcription factors (TF) which play a key role in the regulatory networks. They are mainly known at a structural level but few knowledge is available on their transcriptome behavior (Mitsuda and Ohme-Takagi, 2009).

The objective of this thesis is to study the behavior of TF in stress response in *Arabidopsis thaliana*. The first axis is to identify TF groups who work in response to stress, to characterize and compare them in order to identify typical TF combinations in response to biotic and/or abiotic. Methodological questions considered in this part will be linked to model-based clustering to be able to cluster TF by taking all the available information into account either in the conditional distribution of the TF described by their transcriptomes or in the penalty term when a model has to be selected in a considered collection.

The second axis will be dedicated to the target identification of the TF groups. The framework will be Gaussian graphical models and linear regression models where $n \sim p$. The challenge will be to develop a penalty taking biological knowledge into account and to calibrate it to get powerful inference with a correct FDR control. The subject combines modeling and algorithmic questions and will contribute to a better knowledge of TF of *Arabidopsis thaliana*.

Bibliography

Abdelkarim, M., Vintonenko, N., Starzec, A., Robles, A., Aubert, J., Martin, M.-L., Mourah, S., Podgorniak, M.-P., Rodrigues-Ferreira, S., Nahmias, C., Couraud, P.-O., Doliger, C., Sainte-Catherine, O., Peyri, N., Chen, L., Mariau, J., Etienne, M., Perret, G.-Y., Crepin, M., Poyet, J.-L., Khatib, A.-M., and Di Benedetto, M. (2011). Invading Basement Membrane Matrix Is Sufficient for MDA-MB-231 Breast Cancer Cells to Develop a Stable *In Vivo* Metastatic Phenotype. *PLoS ONE*, **6**(8), e23334.

Andrieu, C. (2003). An Introduction to MCMC for Machine Learning.

Aubourg, S., Martin-Magniette, M.-L., Brunaud, V., Taconnat, L., Bitton, F., Balzergue, S., Jullien, P., Ingouff, M., Thareau, V., Schiex, T., Lecharny, A., and Renou, J. (2007). Analysis of CATMA transcriptome data identifies hundreds of novel functional genes and improves gene models in the Arabidopsis genome. *BMC Genomics*, **8**, 401.

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

Baudry, J., Raftery, A., Celeux, G., Lo, K., and Gottardo, R. (2010). Combining Mixture Components for Clustering. *Journal of Computational and Graphical Statistics*, **9**(2), 332–353.

Baudry, J.-P., Maugis, C., and Michel, B. (2012). Slope Heuristics: overview and implementation. *Statistics and Computing*, **22**, 455–470.

Beal, M. J. and Ghahramani, Z. (2003). The Variational Bayesian EM Algorithm for Incomplete Data: with Application to Scoring Graphical Model Structures. *Bayesian Statistics*, **7**.

Benhamed, M., Martin-Magniette, M.-L., Taconnat, L., Bitton, F., Servet, C., De Clercq, R., De Meyer, B., Buysschaert, C., Rombauts, S., Villarroel, R., Aubourg, S., Beynon, J., Bhalerao, R., Coupland, G., Gruissem, W., Menke, F., Weisshaar, B., Renou, J.-P., Zhou, D.-X., and Hilson, P. (2008). Genome-scale Arabidopsis promoter array identifies targets of the histone acetyltransferase GCN5. *Plant Journal*.

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.

Bérard, C., Martin-Magniette, M.-L., Brunaud, V., Aubourg, S., and Robin, S. (2011). Unsupervised Classification for Tiling Arrays: ChIP-chip and Transcriptome. *Statistical Applications in Genetics and Molecular Biology*, **10**.

Bérard, C., Seifert, M., Mary-Huard, T., and Martin-Magniette, M.-L. (2013). MultiChIPmixHMM: an R package for ChIP-chip data analysis modeling spatial dependencies and multiple replicates. *BMC Bioinformatics*. In revision.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**, 719–725.

Biernacki, C., Celeux, G., Govaert, G., and Langrognet, F. (2006). Model-based cluster and discriminant analysis with the MIXMOD software. *Computational Statistics and Data Analysis*, **51**(2), 587–600.

Bilban, M., Buehler, L., Head, S., Desoye, G., and Quaranta, V. (2002). Defining signal thresholds in DNA microarrays: exemplary application for invasive cancer. *BMC Genomics*, **3**(1), 19.

Bolstad, B., Irizarry, R., strand, M., and Speed, T. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**(2), 185–193.

Bouveyron, C. and Brunet, C. (2012). Simultaneous model-based clustering and visualization in the Fisher discriminative subspace. *Statistics and Computing*, **22**(1), 301–324.

Bouveyron, C., Girard, S., and Schmid, C. (2007). High-Dimensional Data Clustering. *Computational Statistics and Data Analysis*, **52**(1), 502–519.

Brown, C., Goodwin, P., and Sorger, P. (2001). Image metrics in the statistical analysis of DNA microarray data. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(16), 8944–8949.

Brusco, M. J. and Cradit, J. D. (2001). A variable selection heuristic for k-means clustering. *Psychometrika*, **66**(2), 249–270.

Buck, M. and Lieb, J. (2004). ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, **83**(3), 349–360.

Cappé, O., Moulines, E., and Ryden, T. (2005). *Inference in Hidden Markov Models*. Springer Series in Statistics.

Cawley, S., Bekiranov, S., Ng, H. H., Kapranov, P., Sekinger, E. A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A. J., Wheeler, R., Wong, B., Drenkow, J., Yamanaka, M., Patel, S., Brubaker, S., Tammana, H., Helt, G., Struhl, K., and Gingeras, T. R. (2004). Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell*, **116**(4), 499–509.

Celeux, G. and Durand, J.-B. (2008). Selecting hidden markov model state number with cross-validated likelihood. *Computational Statistics*, **23**(4), 541–564.

Celeux, G., Martin-Magniette, M.-L., Maugis, C., and Raftery, A. E. (2011). Witten, d. m., and tibshirani, r. (2010), "a framework for feature selection in clustering, journal of the american statistical association, 105,713-726. *Journal of the American Statistical Association*, **106**(493), 383.

Celton, M., Malpertuy, A., Lelandais, G., and de Brevern, A. (2010). Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments. *BMC Genomics*, **11**(1), 15.

Cenik, C., Derti, A., Mellor, J., Berriz, G., and Roth, F. (2010). Genome-wide functional analysis of human 5' untranslated region introns. *Genome Biology*, **11**(3), R29.

Chadi, S., Young, R., Le Guillou, S., Tilly, G., Bitton, F., Martin-Magniette, M.-L., Soubigou-Taconnat, L., Balzergue, S., Vilotte, M., Peyre, C., Passet, B., Beringue, V., Renou, J.-P., Le Provost, F., Laude, H., and Vilotte, J.-L. (2010). Brain transcriptional stability upon prion protein-encoding gene invalidation in zygotic or adult mouse. *BMC Genomics*, **11**(1), 448.

Chatzis, S. (2010). Hidden Markov Models with Nonelliptically Contoured State Densities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **32**, 2297–2304.

Chiquet, J., Smith, A., Grasseau, G., Matias, C., and Ambroise, C. (2009). SIMoNe: Statistical Inference for MOdular NEtworks. *Bioinformatics*, **25**(3), 417–418.

Cleveland, W. S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, **74**, 829–836.

Cohen, D., Bogeat-Triboulot, M.-B., Tisserant, E., Balzergue, S., Martin-Magniette, M.-L., Lelandais, G., Ningre, N., Renou, J.-P., Tamby, J.-P., Le Thiec, D., and Hummel, I. (2010). Comparative transcriptomics of drought responses in populus: a meta-analysis of genome-wide expression profiling in mature leaves and root apices across two genotypes. *BMC Genomics*, **11**, 630.

Cossegal, M., Chambrier, P., Mbelo, S., Balzergue, S., Martin-Magniette, M.-L., Moing, A., Deborde, C., Guyon, V., Perez, P., and Rogowsky, P. (2008). Transcriptional and Metabolic Adjustments in ADP-Glucose Pyrophosphorylase-Deficient bt2 Maize Kernels. *Plant Physiology*, **146**(4), 1553–1570.

Cubillos, F., Yansouni, J., Khalili, H., Balzergue, S., Elftieh, S., Martin-Magniette, M.-L., Serrand, Y., Lepiniec, L., Baud, S., Dubreucq, B., Renou, J.-P., Camilleri, C., and Loudet, O. (2012). Expression variation in connected recombinant populations of Arabidopsis thaliana highlights distinct transcriptome architectures. *BMC Genomics*, **13**(1), 117.

Dash, M., Choi, K., Scheuermann, P., and Liu, H. (2002). Feature Selection for Clustering - A Filter Solution. *Proceedings of the Second IEEE International Conference on Data Mining*, pages 115–122.

Delmar, P., Robin, S., and Daudin, J. J. (2005). VarMixt: efficient variance modelling for the differential analysis of replicated gene expression data. *Bioinformatics*, **21**(4), 502–508.

Dempster, A., N., L., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Devaney, M. and Ram, A. (1997). Efficient Feature Selection in Conceptual Clustering. *Machine Learning: Proceedings of the Fourteenth International Conference*, pages 92–97.

Dobbin, K., Shih, J. H., and Simon, R. (2003). Statistical design of reverse dye microarrays. *Bioinformatics*, **19**(7), 803–810.

Dobbin, K. K., Shih, J. H., and Simon, R. M. (2005). Comment on "Evaluation of the gene-specific dye bias in cDNA microarray experiments". *Bioinformatics*, **21**(12), 2803–2804.

Dubois, A., Remay, A., Raymond, O., Balzergue, S., Chauvet, A., Maene, M., Pécrix, Y., Yang, S.-H., Jeauffre, J., Thouroude, T., Boltz, V., Martin-Magniette, M.-L., Janczarski, S., Legeai, F., Renou, J.-P., Vergne, P., Le Bris, M., Foucher, F., and Bendahmane, M. (2011). Genomic Approach to Study Floral Development Genes in *Rosa sp. PLoS ONE*, **6**(12), e28455.

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

Efron, B. and Tibshirani, R. (2002). Empirical bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, **23**(1), 70–86.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, **95**(25), 14863–14868.

Elis, S., Batellier, F., Couty, I., Balzergue, S., Martin-Magniette, M.-L., Monget, P., Blesbois, E., and Govoroun, M. (2008). Search for the genes involved in oocyte maturation and early embryo development in the hen. *BMC Genomics*, **9**(1), 110.

Elis, S., Blesbois, E., Couty, I., Balzergue, S., Martin-Magniette, M.-L., and Batellier, F. (2009). Identification of germinal disk region derived genes potentially involved in hen fertility. *Molecular Reproduction and Development*, **76**(11), 1043–1055.

Faivre Rampant, P., Lesur, I., Boussardon, C., Bitton, F., Martin-Magniette, M.-L., Bodenes, C., Le Provost, G., Berges, H., Fluch, S., Kremer, A., and Plomion, C. (2011). Analysis of bac end sequences in oak, a keystone forest tree species, providing insight into the composition of its genome. *BMC Genomics*, **12**(1), 292.

Forster, T., Costa, Y., Roy, D., Cooke, H., and Maratou, K. (2004). Triple-target microarray experiments: a novel experimental strategy. *BMC Genomics*, **5**(1), 13.

Fowlkes, E. B., Gnanadesikan, R., and Kettenring, J. R. (1988). Variable selection in clustering. *Journal of Classification*, **5**(2), 205–228.

Friedman, J. H. and Meulman, J. J. (2004). Clustering objects on subsets of attributes (with discussion). *Journal of the Royal Statistical Society, Series B*, **66**(4), 815–849.

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-Issue), 986–990.

Giraud, C., Huet, S., and Verzelen, N. (2012). Graph Selection with GGMselect. *Statistical Applications in Genetics and Molecular Biology*, **11**(3).

Gollery, M., Harper, J., Cushman, J., Mittler, T., Girke, T., Zhu, J.-K., Bailey-Serres, J., and Mittler, R. (2006). What makes species unique? the contribution of proteins with obscure features. *Genome Biology*, **7**(7), R57.

Grennan, A. K. (2006). Genevestigator. Facilitating Web-Based Gene-Expression Analysis. *Plant Physiology*, **141**(4), 1164–1166.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, **3**, 1157–1182.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, New York, Second Edition.

Hennig, C. (2010). Methods for merging gaussian mixture components. *Advances in Data Analysis and Classification*, pages 3–34.

Herbette, S., Taconnat, L., Hugouvieux, V., Martin-Magniette, M.-L., Cuine, S., Auroy, P., Richaud, P., Forestier, C., Bourguignon, J., Renou, J.-P., Vavasseur, A., and Leonhart, N. (2006). Genome-wide transcriptome profiling of the early cadmium response of Arabidopsis roots and shoots. *Biochimie*, **11**, 1751–1765.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian Model Averaging: A Tutorial. *Statistical science*, **14**(4), 382—417.

Horan, K., Jang, C., Bailey-Serres, J., Mittler, R., Shelton, C., Harper, J. F., Zhu, J.-K., Cushman, J. C., Gollery, M., and Girke, T. (2008). Annotating genes of known and unknown function by large-scale coexpression analysis. *Plant Physiology*, **147**(1), 41–57.

Houel, C., Martin-Magniette, M.-L., Nicolas, S., Lacombe, T., Le Cunff, L., Franck, D., Torregrosa, L., Conéjéro, G., Lalet, S., This, P., and Adam-Blondon, A.-F. (2013). Genetic variability of berry size in the grapevine (Vitis viniferaL.). *Australian Journal of Grape and Wine Research*, **19**(2), 208–220.

Huber, P. J. (1981). *Robust Statistics*. John Wiley & Sons, New York.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, **31**(4), e15.

Jaakkola, T. S. and Jordan, M. I. (1998). Improving the mean field approximation via the use of mixture distributions. In *Proceedings of the NATO Advanced Study Institute on Learning in graphical models*, pages 163–173, Erice, Italy. Kluwer Academic Publishers.

Jammes, F., Lecomte, P., de Almeida-Engler, J., Bitton, F., Martin-Magniette, M.-L., Renou, J.-P., Abad, P., and Favery, B. (2005). Genome-wide expression profiling of the host response to root-knot nematode infection in Arabidopsis. *Plant Journal*, **44**(3), 447–458.

Ji, H. and Wong, W. (2005). TileMap: create chromosomal map of tiling array hybridizations. *Bioinformatics*, **21**(18), 3629–3636.

Jiang, D., Tang, C., and Zhang, A. (2004). Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, **16**(11), 1370–1386.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1994). *Continuous Univariate Distributions. (2nd Edition)*. Wiley Series in Probability and Statistics. John Wiley & Sons, New York.

Jouve, P.-E. and Nicoloyannis, N. (2005). A filter feature selection method for clustering. *Proceedings of International symposium on Methodologies for Intelligent Systems*, pages 583–593.

Keles, S. (2007). Mixture Modeling for Genome-Wide Localization of Transcription Factors. *Biometrics*, **63**(1), 10–21.

Kelley, R., Feizi, H., and Ideker, T. (2008). Correcting for gene-specific dye bias in DNA microarrays using the method of maximum likelihood. *Bioinformatics*, **24**(1), 71–77.

Kerr, M., Afshari, C., Bennett, L., Bushel, P., Martinez, J., Walker, N., and Churchill, G. (2002). Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica*, **12**, 203–217.

Kim, S., Tadesse, M. G., and Vannucci, M. (2006). Variable selection in clustering via Dirichlet process mixture models. *Biometrika*, **93**(4), 877–893.

Kohavi, R. and John, G. H. (1997). Wrappers for Feature Subset Selection. *Artificial Intelligence*, **97**(1-2), 273–324.

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D. L., Garcia-Hernandez, M., Karthikeyan, A. S., Lee, C. H., Nelson, W. D., Ploetz, L., Singh, S., Wensel, A., and Huala, E. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research*, **40**(D1), D1202–D1210.

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.

Li, J. (2005). Clustering based on a multilayer mixture model. *Journal of Computational and Graphical Statistics*, **14**(3), 547–568.

Li, W., Meyer, C., and Liu, X. (2005). A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics*, **21**(suppl 1), i274–i282.

Lurin, C., Andréas, C., Aubourg, S., Bellaoui, M., Bitton, F., Bruyère, C., Caboche, M., Debast, C. Gualberto, J., Hoffmann, B., Lecharny, A. Le Ret, M., Martin-Magniette, M.-L., Mireau, H., Peeters, N., Renou, J.-P., Szurek, B., Taconnat, L., and Small, I. (2004). Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. *Plant Cell*, **16**(8), 2089–2103.

Madigan, D. and Hutchinson, F. (1995). Enhancing the Predictive Performance of Bayesian Graphical Models. *Communications in statistics: Theory and methods*, **24**, 2271–2292.

Madigan, D., Raftery, A., Wermuth, N., York, J., and Zucchini, W. (1994). Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. *Journal of the American Statistical Association*, **89**, 1535–1546.

Margaritis, T., Lijnzaad, P., van Leenen, D., Bouwmeester, D., Kemmeren, P., van Hooff, S. R., and Holstege, F. C. (2009). Adaptable gene-specific dye bias correction for two-channel DNA microarrays. *Molecular Systems Biology*, **5**.

Marin, J. and Robert, C. P. (2009). Importance sampling methods for Bayesian discrimination between embedded models. *ArXiv e-prints*.

Mary-Huard, T., Daudin, J.-J., Robin, S., Bitton, F., Cabannes, E., and Hilson, P. (2004). Spotting effect in microarray experiments. *BMC Bioinformatics*, **5**.

Mary-Huard, T., Perduca, V., Martin-Magniette, M.-L., and Blanchard, G. (2013). Error rate control for classification rules in multi-class mixture models. In *45 ième Journées de Statistiques*, page 106. Société Française de Statistiques.

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., Martin-Magniette, M.-L., Tamby, J.-P., Renou, J.-P., Lecharny, A., Aubourg, S., and Celeux, G. (2009a). Sélection de variables pour la classification par mélanges gaussiens pour prédire la fonction des gènes orphelins. *Modulad*, **40**.

Maugis, C., Celeux, G., and Martin-Magniette, M.-L. (2009b). Variable Selection for Clustering with Gaussian Mixture Models. *Biometrics*, **65**, 701–709.

Maugis, C., Celeux, G., and Martin-Magniette, M.-L. (2009c). Variable selection in model-based clustering: A general variable role modeling. *Computational Statistics and Data Analysis*, **53**, 3872–3882.

Maugis, C., Celeux, G., and Martin-Magniette, M.-L. (2011). Variable selection in model-based discriminant analysis. *Journal of Multivariate Analysis*, **102**(10), 1374–1387.

Maugis-Rabusseau, C., Martin-Magniette, M.-L., and Pelletier, S. (2012). SelvarClustMV: Variable selection approach in model-based clustering allowing for missing values. *Journal de la Société Française de Statistiques*, **153**(2).

McLachlan, G. (1992). *Discriminant Analysis and Statistical Pattern Analysis*. Wiley-Interscience, New York.

McLachlan, G., Bean, R., and Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**(3), 413–422.

McLachlan, G. J. and Peel, D. (2000). *Finite mixture models*. Wiley Series in Probability and Statistics. John Wiley, New York.

Mcnicholas, P. D. and Murphy, T. B. (2008). Parsimonious Gaussian mixture models. *Statistics and Computing*, **18**(3), 285–296.

Meynet, C. and Maugis-Rabusseau, C. (2012). A sparse variable selection procedure in model-based clustering. Technical Report 00734316, Hal.

Miller, A. J. (1990). *Subset Selection in Regression*. Chapman and Hall.

Mitsuda, N. and Ohme-Takagi, M. (2009). Functional Analysis of Transcription Factors in Arabidopsis. *Plant and Cell Physiology*, **50**(7), 1232–1248.

Moghaddam, A., Roudier, F., Seifert, M., Bérard, C., Martin-Magniette, M., Ashtiyani, R., Houben, A., Colot, V., and Mette, M. (2011). Additive inheritance of histone modifications in arabidopsis thaliana intra-specific hybrids. *The Plant Journal*.

Moreau, M., Azzopardi, M., Clément, G., Dobrenel, T., Marchive, C., Renne, C., Martin-Magniette, M., Taconnat, L., Renou, J., Robaglia, C., and Meyer, C. (2012). Mutations in the Arabidopsis Homolog of LST8/GL, a Partner of the Target of Rapamycin Kinase, Impair Plant Growth, Flowering, and Metabolic Adaptation to Long Days. *The Plant Cell Online*.

Murphy, B. T., Raftery, A. E., and Dean, N. (2010). Variable Selection and Updating in Model-Based Discriminant Analysis for High-Dimensional Data with Food Authenticity Applications. *Annals of Applied Statistics*, **4**(1), 396–421.

Navajas, M., Migeon, A., Alaux, C., Martin-Magniette, M., Robinson, G., Evans, J., Cros-Arteil, S., Crauser, D., and Le Conte, Y. (2008). Differential gene expression of the honey bee Apis mellifera associated with Varroa destructor infection. *BMC Genomics*, **9**(1), 301.

Papastamoulis, P., Martin-Magniette, M.-L., and Maugis-Rabusseau, C. (2013). Efficient estimation of mixtures of Poisson Generalized Linear Models with large number of components. *submitted*.

Persson, S., Wei, H., Milne, J., Page, G., and Somerville, C. (2005). Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets.

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature Genetics*, **32**, 496–501.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77**, 257–286.

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

Raftery, A. E., Hoeting, J. A., and Madigan, D. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, **92**, 179–191.

Raftery, A. E., Zheng, Y., N-We, Clyde, M., Hoeting, J., and Madigan, D. (2003). Long-Run Performance of Bayesian Model Averaging. *Journal of the American Statistical Association*, **98**, 931–938.

Ramel, F., Sulmon, C., Cabello-Hurtado, F., Taconnat, L., Martin-Magniette, M.-L., Renou, J.-P., El Amrani, A., Couee, I., and Gouesbet, G. (2007). Genome-wide interacting effects of sucrose and herbicide-mediated stress in Arabidopsis thaliana: novel insights into atrazine toxicity and sucrose-induced tolerance. *BMC Genomics*, **8**(1), 450.

Rau, A., Celeux, G., Martin-Magniette, M.-L., and Maugis-Rabusseau, C. (2011). Clustering high-throughput sequencing data with Poisson mixture models. Technical Report RR-7786, INRIA.

Rengel, D., Arribat, S., Maury, P., Martin-Magniette, M.-L., Hourlier, T., Laporte, M., Vars, D., Carrre, S., Grieu, P., Balzergue, S., Gouzy, J., Vincourt, P., and Langlade, N. B. (2012). A Gene-Phenotype Network Based on Genetic Variability for Drought Responses Reveals Key Physiological Processes in Controlled and Natural Environments. *PLoS ONE*, **7**(10), e45249.

Reymond, M. C., Brunoud, G., Chauvet, A., Martnez-Garcia, J. F., Martin-Magniette, M.-L., Monéger, F., and Scutt, C. P. (2012). A Light-Regulated Genetic Module Was Recruited to Carpel Development in Arabidopsis following a Structural Change to SPATULA. *The Plant Cell Online*.

Roudier, F., Ahmed, I., Berard, C., Sarazin, A., Mary-Huard, T., Cortijo, S., Bouyer, D., Caillieux, E., Duvernois-Berthet, E., Al-Shikhley, L., Giraut, L., Despres, B., Drevensek, S., Barneche, F., Derozier, S., Brunaud, V., Aubourg, S., Schnittger, A., Bowler, C., Martin-Magniette, M.-L., Robin, S., Caboche, M., and Colot, V. (2011). Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. *EMBO Journal*, **30**, 1928–1938.

Ruffel, S., Freixes, S., Balzergue, S., Tillard, P., Jeudy, C., Martin-Magniette, M.-L., van der Merwe, M. J., Kakar, K., Gouzy, J., Fernie, A. R., Udvardi, M., Salon, C., Gojon, A., and Lepetit, M. (2008). Systemic Signaling of the Plant Nitrogen Status Triggers Specific Transcriptome Responses Depending on the Nitrogen Source in Medicago truncatula. *Plant Physiology*, **146**(4), 2020–2035.

Rustici, G., Kolesnikov, N., Brandizi, M., Burdett, T., Dylag, M., Emam, I., Farne, A., Hastings, E., Ison, J., Keays, M., Kurbatova, N., Malone, J., Mani, R., Mupo, A., Pedro Pereira, R., Pilicheva, E., Rung, J., Sharma, A., Tang, Y. A., Ternent, T., Tikhonov, A., Welter, D., Williams, E., Brazma, A., Parkinson, H., and Sarkans, U. (2013). ArrayExpress updatetrends in database growth and links to data analysis tools. *Nucleic Acids Research*, **41**(D1), D987–D990.

Schiex, T., Moisan, A., and Rouz, P. (2001). Eugène: An Eukaryotic Gene Finder That Combines Several Sources of Evidence. In O. Gascuel and M.-F. Sagot, editors, *Computational Biology*, volume 2066 of *Lecture Notes in Computer Science*, pages 111–125. Springer Berlin Heidelberg.

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

Sharan, R., Elkon, R., and Shamir, R. (2002). Cluster analysis and its applications to gene expression data. In *Ernst Schering Workshop on Bioinformatics and Genome Analysis*. Springer Verlag.

Si, Y., Liu, P., Li, P., and Brutnell, T. (2011). Model-based clustering for RNA-seq data. Technical report, Iowa State University.

Staal, Y., van Herwijnen, M., van Schooten, F., and van Delft, J. (2005). Application of four dyes in gene expression analyses by microarrays. *BMC Genomics*, **6**(1), 101.

Stolc, V., Gauhar, Z., Mason, C., Halasz, G., van Batenburg, M., Rifkin, S., Hua, S., Herreman, T., Tongprasit, W., Barbano, P., Bussemaker, H., and White, K. (2004). A gene expression map for the euchromatic genome of Drosophila melanogaster. *Science*, **306**, 655–660.

Sun, W. and Cai, T. (2009). Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society*, **71**, 393–424.

Tadesse, M. G., Sha, N., and Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, **100**(470), 602–617.

Teixeira, V. H., Olaso, R., Martin-Magniette, M.-L., Lasbleiz, S., Jacq, L., Oliveira, C. R., Hilliquin, P., Gut, I., Cornelis, F., and Petit-Teixeira, E. (2009). Transcriptome Analysis Describing New Immunity and Defense Genes in Peripheral Blood Mononuclear Cells of Rheumatoid Arthritis Patients. *PLoS ONE*, **4**(8), e6803.

Turck, F., Roudier, F., Farrona, S., Martin-Magniette, M.-L., Guillaume, E., Buisine, N., Gagnot, S., Martienssen, R. A., Coupland, G., and Colot, V. (2007). Arabidopsis TFL2/LHP1 Specifically Associates with Genes Marked by Trimethylation of Histone H3 Lysine 27. *PLoS Genet*, **3**(6), e86.

Martin-Magniette, M.-L. and Robin, S. (2004). *Informatique pour l'analyse du transcriptome*, chapter Techniques statistique pour l'analyse du transcriptome. Herms, Boulicaut, J.-F. and Gandrillon, O. edition.

Martin-Magniette, M.-L., Aubert, J., Cabannes, E., and Daudin, J.-J. (2005a). Answer to the comments of K. Dobbin, J. Shih and R. Simon on the paper "Evaluation of the gene-specific dye-bias in cDNA microarray experiments". *Bioinformatics*, **21**(14), 3065–3065.

Martin-Magniette, M.-L., Aubert, J., Cabannes, E., and Daudin, J.-J. (2005b). Evaluation of the gene-specific dye bias in cDNA microarray experiments. *Bioinformatics*, **21**(9), 1995–2000.

Martin-Magniette, M.-L., Mary-Huard, T., Bérard, C., and Robin, S. (2008a). ChIPmix: mixture model of regressions for two-color ChIP-chip analysis. *Bioinformatics*, **24**(16), i181–186.

Martin-Magniette, M.-L., Aubert, J., Bar-Hen, A., Elftieh, S., Magniette, F., Renou, J.-P., and Daudin, J.-J. (2008b). Normalization for triple-target microarray experiments. *BMC Bioinformatics*, **9**(1), 216.

Vandel, J., Mangin, B., Vignes, M., Leroux, D., Loudet, O., Martin-Magniette, M.-L., and De Givry, S. (2012). Inférence de réseaux de régulation de gnes au travers de scores étendus dans les réseaux bayésiens. *Revue d'Intelligence Artificielle*, **26**(6), 679–708.

Verzelen, N. (2012). Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics*, **6**, 38–90.

Volant, S., Martin-Magniette, M.-L., and Robin, S. (2012). Variational Bayes approach for model aggregation in unsupervised classification with Markovian dependency. *Computational Statistics and Data Analysis*, **56**(8), 2375–2387.

Volant, S., Bérard, C., Martin-Magniette, M.-L., and Robin, S. (2013). Hidden markov models with mixtures as emission distributions. *Statistics and Computing*, pages 1–12.

Volinsky, C. T., Madigan, D., Raftery, A. E., and Kronmal, R. A. (1997). Bayesian Model Averaging in proportional hazard models: Assessing the risk of a stroke. *Journal of the Royal Statistical Society, Series C*, **46**(4), 433–448.

Wang, J., Lee, J., and Zhang, C. (2003). Kernel Trick Embedded Gaussian Mixture Model. In R. Gavald, K. Jantke, and E. Takimoto, editors, *Algorithmic Learning Theory*, volume 2842 of *Lecture Notes in Computer Science*, pages 159–174. Springer Berlin Heidelberg.

Weber, M., Hellmann, I., Stadler, M. B., Ramos, L., Pääbo, S., Rebhan, M., and Schübeler, D. (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nature Genetics*, **39**, 457 – 466.

Wu, Z., Irizarry, R. A., Gentleman, R., Martinez-Murillo, F., and Spencer, F. (2004). A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association*, **99**(468), 909–917.

Yang, Y., Dudoit, S., Luu, P., Lin, D., Peng, V., Ngai, J., and Speed, T. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, **30**.

Yang, Y. H. and Thorne, N. (2003). Single channel normalisation for cDNA microarray data. *IMS Lecture Notes– Monograph Series*, **40**, 403–418.

Young, D. M. and Odell, P. L. (1986). Feature-subset selection for statistical classification problems involving unequal covariance matrices. *Communication in Statistics-Theory and Methods*, **15**, 137–157.

Zhang, Q. and Wang, H. (2008). A BIC Criterion for Gaussian Mixture Model Selection with Application in Discriminant Analysis. Technical report, Guanghua School of Management, Peking University.

Zhang, X., Clarenz, O., Cokus, S., Bernatavichute, Y. V., Pellegrini, M., Goodrich, J., and Jacobsen, S. E. (2007). Whole-genome analysis of histone h3 lysine 27 trimethylation in *Arabidopsis*. *PLoS Biology*, **5**(5), e129.