Mélanges gaussiens bidimensionnels pour la comparaison de deux échantillons de chromatine immunoprécipitée

Caroline Bérard[1], Marie-Laure Martin-Magniette[1,2], Alexandra To[3], François Roudier[3], Vincent Colot[3] et Stéphane Robin[1].

[1] UMR AgroParisTech/INRA MIA 518, 16 rue Claude Bernard, PARIS Cedex 05.
[2] UMR INRA 1165 - CNRS 8114 - UEVE URGV, 2 rue Gaston Crémieux, EVRY.
[3] UMR CNRS 8186, Département de Biologie, 46 rue d'Ulm, PARIS Cedex 05.

caroline.berard@agroparistech.fr , marie_laure.martin@agroparistech.fr
to@biologie.ens.fr , roudier@biologie.ens.fr
colot@biologie.ens.fr , stephane.robin@agroparistech.fr

Résumé L'immunoprécipitation de la chromatine (ChIP) permet d'étudier les interactions entre les protéines et l'ADN ainsi que différents états chromatiniens. Le ChIP-chip est une technique combinant l'immunoprécipitation de la chromatine avec le principe des puces à ADN, ce qui permet une étude à l'échelle du génome. Nous nous intéressons ici à l'analyse des différences entre deux échantillons d'ADN immunoprécipité. Biologiquement, on s'attend à distinguer quatre groupes différents : un groupe d'ADN non-immunoprécipité, un groupe d'ADN immunoprécipité identiquement dans les deux échantillons et deux groupes dans lesquels l'ADN est immunoprécipité en quantités différentes. Nous modélisons ces données par un mélange de gaussiennes bidimensionnelles à quatre composants. Les matrices de variance sont contraintes afin d'intégrer des connaissances biologiques. Les paramètres sont estimés par l'algorithme EM. Nous appliquons cette méthode pour étudier la différence de méthylation d'une histone entre l'écotype sauvage de la plante modèle *Arabidopsis thaliana* et un mutant.
Mots-clés : Mélange gaussien, décomposition spectrale, algorithme EM, ChIP-chip.

Résumé Chromatin immunoprecipitation (ChIP) enables to investigate interactions between proteins and DNA and also various chromatin states. ChIP-chip is a well-established procedure combining chromatin immunoprecipitation with DNA microarrays, which allows a study of the whole genome. We are interested in the analyze of the differences between two immunoprecipitated DNA samples. From a biological point of view, we expect to distinguish four different groups : a group of non-immunoprecipited DNA, a group of immunoprecipited DNA in both samples, and then two groups in which DNA is differently immunoprecipited. We propose to model these data with a mixture of two-dimensional Gaussians with four components. Biological knowledges are included as constraints on the variance matrices. The parameters are estimated by the EM algorithm. This method is applied to NimbleGen data in order to study the histone methylation difference between the wild ecotype of the model plant *Arabidopsis thaliana* and a mutant.
Keywords : Gaussian mixture, eigenvalue decomposition, EM algorithm, ChIP-chip.

1 Introduction

La connaissance des mécanismes de régulation des gènes est essentielle pour comprendre certains concepts biologiques importants. On sait par exemple que le développement d'un organisme dépend grandement de l'harmonisation de l'expression de ses gènes. Après le séquençage entier des génomes à grande échelle, le défi consiste donc aujourd'hui à comprendre le fonctionnement des gènes, c'est-à-dire à déterminer leur fonction et leur patron d'expression.

Dans le noyau des cellules eucaryotes, l'ADN est fractionné en chromosomes et il est condensé sous forme de chromatine. La chromatine est un complexe ADN-protéines qui joue un rôle essentiel dans le contrôle de l'activité des gènes. Les protéines présentes sont principalement des histones. La condensation de l'ADN en chromatine s'organise de manière séquentielle et ordonnée. En premier lieu, 147 paires de bases d'ADN s'enroulent autour d'un octamère d'histones pour former un nucléosome. Dans un second niveau d'organisation, les nucléosomes se compactent et forment une hélice. Cette hélice est finalement condensée en euchromatine (condensation légère) ou en hétérochromatine (condensation plus prononcée) constituant un troisième niveau d'organisation. Les gènes localisés dans l'euchromatine peuvent être plus facilement transcrits car la condensation est légère. Cette structure d'organisation du génome dans le noyau constitue en elle-même un mécanisme de répression ou d'activation de la transcription des gènes. En effet, pour activer la transcription d'un gène donné dans une cellule, la chromatine comprise dans la région de contrôle du gène doit être modifiée ou altérée de façon à être permissive à la transcription. Les modifications post-traductionnelles d'histone (comme la méthylation, l'acétylation, l'ubiquitination ou la phosphorilation) sont des mécanismes impliqués dans la régulation de l'expression des gènes (Turck *et al.* [18]).

L'immunoprécipitation de la chromatine (ChIP) permet d'étudier les interactions entre les protéines et l'ADN ainsi que différents états chromatiniens associés à des états d'activité distincts du génome. Le ChIP-chip est une technique combinant l'immunoprécipitation de la chromatine avec le principe des puces à ADN (Amaratunga et Cabrera [1]), ce qui permet une étude à l'échelle du génome. Habituellement dans une expérience de ChIP-chip, les deux échantillons co-hybridés sont les fragments d'ADN associés à la protéine d'intérêt ou à une marque chromatinienne (IP) et l'ADN génomique total (INPUT). Le but est ensuite de détecter les sondes de la puce pour lesquelles il y a un signal IP afin d'identifier les régions génomiques où la protéine d'intérêt se fixe.

Buck et Lieb [7] ont montré la nécessité de développer de nouvelles méthodes statistiques pour détecter les sondes enrichies dans les expériences de ChIP-chip. Récemment, deux stratégies ont été largement appliquées : la première tient compte de la structure spatiale des données (Cawley *et al.* [9], Keles [14]) et la seconde considère que la totalité des sondes peut être divisée en deux populations : les sondes enrichies et les non-enrichies (Buck et Lieb [7], Turck *et al.* [18], Martin-Magniette *et al.* [15]). Différentes méthodes statistiques ont été proposées pour distinguer ces deux populations : toutes sont fondées sur la distribution du log-ratio $\log(IP/INPUT)$ (Buck et Lieb [7], Turck *et al.* [18]), exceptée la méthode proposée par Martin-Magniette *et al.* [15] qui utilise un mélange de régressions pour modéliser la loi de l'IP conditionnellement à l'INPUT.

La technique du ChIP-chip permet également d'étudier directement la différence entre deux échantillons d'ADN immunoprécipités, sans hybrider sur la puce l'ADN génomique total (INPUT). À notre connaissance il n'existe pas de méthode pour analyser ce type de données (IP/IP) dans la littérature. Les méthodes de segmentation initialement développées pour l'analyse des données CGH (Hupé *et al.* [13], Olshen *et al.* [16], Picard *et al.* [17]) pourraient être utilisées, mais les régions génomiques non immunoprécipitées et les régions immunoprécipitées identiquement dans les deux échantillons seraient indistinguables. De plus ces méthodes sont assez coûteuses en temps de calcul pour des puces tiling-array qui ont un grand nombre de sondes.

L'objectif de notre travail est de proposer une modélisation conjointe des signaux IP obtenus par un modèle de mélange de gaussiennes bi-dimensionnelles. La description des données est détaillée section 2. Les mélanges gaussiens bidimensionnels modélisés à l'aide d'une décomposition de la matrice de variance sont étudiés section 3. Les connaissances biologiques sont prises en compte sous forme de contraintes sur les paramètres du modèle et sur le nombre de composants. Cette modélisation est détaillée dans la section 4. Une application de la méthode sur des données issues de la technologie NimbleGen est présentée dans la section 5.

2 Description des données

Les données analysées concernent la plante modèle *Arabidopsis thaliana*. Les deux échantillons co-hybridés sur la puce visent à étudier le comportement de l'histone H3 diméthylée au niveau de la lysine 9 (H3K9me2). On compare un échantillon sauvage et un échantillon mutant (mutant nrpdlalb).
L'expérience est faite en dye-swap (Boulicaut et Gandrillon [6]) : le principe est de faire une répétition technique en inversant les marquages. Chaque traitement est ainsi marqué par les deux fluorochromes, ce qui permet de contrôler le biais dû au marquage (biais technique). Les intensités des signaux sont ensuite moyennées sur le dye-swap.
La puce à ADN utilisée est une puce tiling-array à oligos courts issue de la technologie NimbleGen. Cette puce permet d'étudier le génome nucléaire d'*Arabidopsis thaliana*, composé de cinq chromosomes et des génomes mitochondrial et chloroplastique. La puce est constituée d'environ 700 000 sondes.

Lorsque l'on étudie des données de ChIP-chip IP/IP, on s'attend à distinguer quatre groupes (cf Figure 1) :

- Un groupe d'intensité faible qui correspond aux séquences d'ADN qui ne sont pas immunoprécipitées (bruit).
- Un groupe où les séquences d'ADN sont immunoprécipitées en même quantité chez le sauvage et chez le mutant. Cela correspond aux endroits sur le génome où l'histone est méthylée identiquement dans les deux échantillons. Ce groupe sera défini dans la suite comme groupe normal.
- Deux groupes où les séquences d'ADN sont immunoprécipitées en quantités différentes

chez le sauvage et chez le mutant. Le taux de méthylation de l'histone H3K9me2 peut être plus faible chez le mutant (groupe appauvri), ou bien au contraire, plus élevé (groupe enrichi).

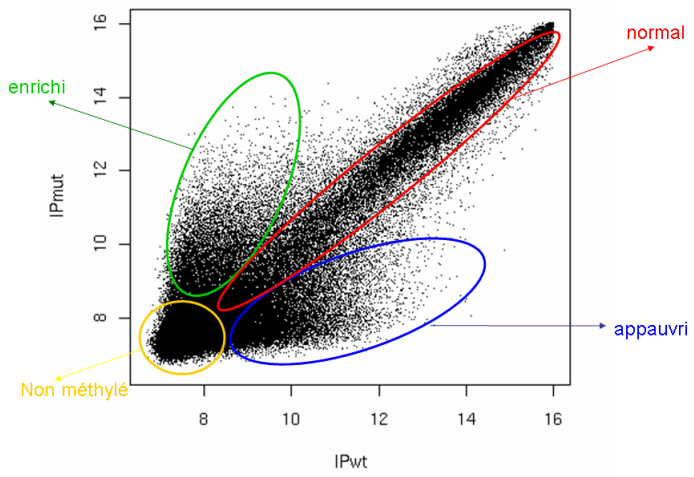

FIG. 1 – Comparaison de deux échantillons de chromatine immunoprécipitée (sauvage vs mutant) : Identification schématique des différents groupes.

3 Modèle de mélanges gaussiens bidimensionnels

Dans cette section, nous rappelons brièvement l'approche de classification par les mélanges gaussiens et reprenons la modélisation des modèles de mélanges gaussiens à l'aide d'une décomposition des matrices de variance, puis nous appliquons certains modèles définis dans Biernacki *et al.* [4] à nos données.

3.1 Approche par classification

Si le but de l'analyse est la classification, le label de chaque donnée est manquant au regard de l'échantillon observé. Notons Z_{ik}, ce label pour l'individu i, qui est une variable aléatoire égale à 1 si le point x_i appartient à la population k et 0 sinon. Les variables $\{Z_1, ...Z_n\}$ (avec $Z_i = \{Z_{i1}, ..., Z_{iK}\}$) sont supposées indépendantes et suivent une loi multinomiale de probabilités $\pi_1, ..., \pi_K$, qui sont les proportions des K classes dans le mélange. Si nous notons Y le vecteur des données complètes (X, Z) où seul X est observé, alors cette reformulation montre clairement que les modèles de mélange peuvent être vus comme un cas particulier des modèles à structure cachée comme par exemple les modèles de Markov cachés (Cappé *et al.* [8], Ephraim et Merhav [12]), la différence étant que les variables $\{Z_1, ...Z_n\}$ sont supposées ici indépendantes.

Dans notre travail, la variable observée $X_i = (X_{1i}, X_{2i})$ est le signal log-IP de chaque échantillon pour la sonde i et nous supposons que les observations proviennent d'un mélange de densités gaussiennes. La densité du couple s'écrit :

$$f(X_i, \psi) = \sum_{k=1}^{K} \pi_k \phi(X_i|\mu_k, \Sigma_k),$$

où π_k est la proportion du k-ième composant du mélange ($0 < \pi_k < 1 \ \forall k = 1, ..., K$ et $\sum_{k=1}^{K} \pi_k = 1$), $\psi = (\pi_1, ..., \pi_{K-1}, \mu_1, ..., \mu_K, \Sigma_1, ..., \Sigma_K)$ est le vecteur des paramètres du mélange et $\phi(.|\mu_k, \Sigma_k)$ est la densité d'une distribution gaussienne bidimensionnelle de moyenne μ_k et de variance Σ_k définis au point x_i par :

$$\phi(x_i|\mu_k, \Sigma_k) = \frac{1}{2\pi}[\det(\Sigma_k)]^{-1/2} \exp\left\{-\frac{1}{2}(x_i - \mu_k)'\Sigma_k^{-1}(x_i - \mu_k)\right\},$$

où M' représente la transposée de M.

Nous calculons les probabilités conditionnelles que la sonde i appartienne à chacun des groupes sachant l'ensemble des observations. Nous rappelons que par définition, la probabilité conditionnelle que la sonde i appartienne au groupe k sachant l'ensemble des observations est définie par :

$$\tau_{ik} = \frac{\hat{\pi}_k \phi(X_i|\hat{\mu}_k, \hat{\Sigma}_k)}{\sum_{l=1}^{K} \hat{\pi}_l \phi(X_i|\hat{\mu}_l, \hat{\Sigma}_l)}.$$

Nous pouvons ensuite classer la sonde i en l'attribuant au groupe pour lequel la probabilité conditionnelle est la plus grande (règle du Maximum A Posteriori).

3.2 Paramétrisation spectrale des matrices de variance

La densité gaussienne modélise une distribution ellipsoïdale de centre μ_k dont les caractéristiques géométriques (volume, forme, orientation) sont définies à l'aide d'une décomposition spectrale de la matrice de variance Σ_k. Pour cela, nous reprenons une paramétrisation proposée par Banfield et Raftery [2] qui permet de proposer de nombreux modèles de classification. Cette paramétrisation considère la décomposition spectrale des matrices de variance :

$$\Sigma_k = \lambda_k D_k A_k D_k' \ , \tag{1}$$

où λ_k représente le volume ($\lambda_k = \det(\Sigma_k)^{1/2}$), D_k représente l'orientation et A_k représente la forme de l'ellipse. La matrice D_k est la matrice des vecteurs propres de Σ_k et A_k est une matrice diagonale telle que $\det(A_k) = 1$ avec les valeurs propres normalisées de Σ_k sur la diagonale dans l'ordre décroissant. En permettant aux paramètres volumes, formes et orientations de varier ou d'être égaux entre les classes, on obtient 14 modèles de mélanges gaussiens différents et facilement interprétables. Les 14 modèles sont détaillés dans Celeux et Govaert [10] : il y a 8 modèles généraux, 4 modèles avec des matrices de variance diagonales et 2 modèles avec des formes sphériques ($A_k = I$).

3.3 Application de 4 modèles de classification aux données

Les 14 modèles de classification (Celeux et Govaert [10]) sont implémentés dans le logiciel MIXMOD [4]. À la vue des données IP/IP (cf Figure 1), nous considérons uniquement les modèles à quatre composants d'orientations différentes, c'est-à-dire les modèles $\lambda D_k A D'_k$, $\lambda_k D_k A D'_k$, $\lambda D_k A_k D'_k$ et $\lambda_k D_k A_k D'_k$. En reprenant les conventions de Celeux et Govaert [10], nous notons λ (respectivement D, A) lorsque le volume (respectivement l'orientation, la forme) est égal pour tous les composants, et λ_k (respectivement D_k, A_k) lorsque le volume (respectivement l'orientation, la forme) est différent pour tous les composants.
On peut choisir le meilleur modèle à l'aide du critère BIC ou du critère ICL. Le critère BIC (Bayesian Information Criterion, Schwarz (1978)) est très utilisé pour les modèles à structure cachée, en particulier les modèles de mélange. Soit $x = (x_1, ..., x_n)$ un n-échantillon où $x_i = (x_{i1}, x_{i2})$ est le signal log-IP observé pour un individu i, le critère BIC du modèle m vaut :

$$BIC_m = -2\log\left\{f(x|\hat{\psi}_m)\right\} + \nu_m \log(n),$$

où $\hat{\psi}_m$ est l'estimateur des paramètres pour le modèle m et ν_m est le nombre de paramètres du modèle m. Le critère ICL (Integrated Complete-data Likelihood, Biernacki *et al.* [5]) prend en compte la capacité d'un modèle de mélange à révéler une structure en classes dans les données. Il correspond au critère BIC pénalisé par un terme d'entropie qui mesure le degré d'imbrication des composants :

$$ICL_m = BIC_m + H_m,$$

où H_m correpond à l'entropie du modèle m, avec :

$$H_m = -2\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log(\tau_{ik}).$$

Les deux critères sélectionnent le modèle $\lambda_k D_k A_k D'_k$. Ce modèle est celui qui a le plus de paramètres à estimer (23 paramètres pour un mélange de 4 gaussiennes bidimensionnelles), ce qui n'est pas un problème étant donné le très grand nombre de données (environ 150 000 observations par jeu de données).
Les résultats obtenus avec le modèle $\lambda_k D_k A_k D'_k$ ne nous satisfont pas (cf Figure 2). En effet, un seul composant couvre les groupes enrichi et appauvri et trois composants sont presque concentriques autour du groupe d'ADN non immunoprécipité (bruit). Ceci est dû au fait que la densité de points est beaucoup plus importante au niveau du groupe d'ADN non immunoprécipité qui regroupe environ 50% des données.

Les modèles non choisis par les critères BIC et ICL et qui considèrent un volume, λ, constant pour les quatres composants sont un peu meilleurs du point de vue de l'interprétation, mais deux composants sont très chevauchants et on ne retrouve pas le groupe d'ADN non immunoprécipité. Beaucoup de sondes sont alors classées dans le groupe appauvri à tort (cf Figure 3).

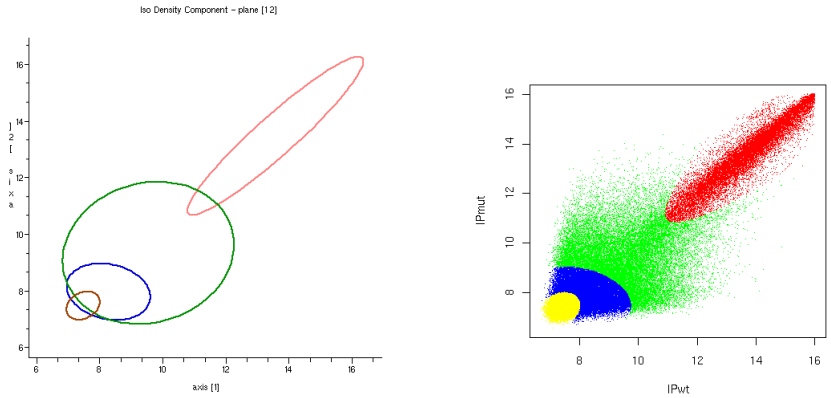

FIG. 2 – Droite : Isodensité des 4 gaussiennes pour le modèle $\lambda_k D_k A_k D'_k$, Gauche : Classement des sondes en 4 groupes avec la règle du MAP

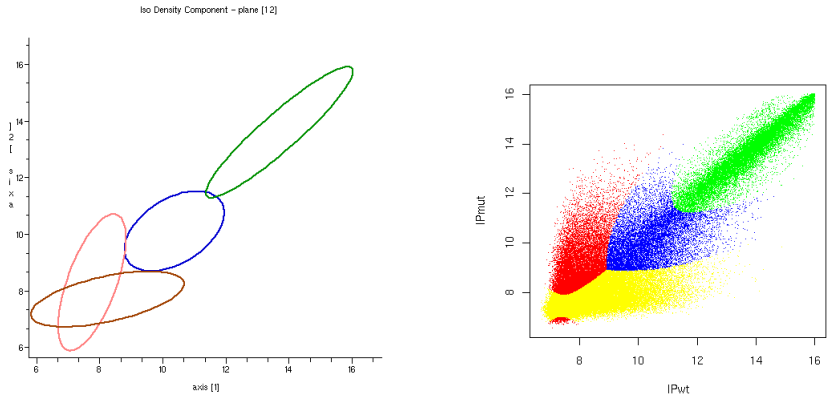

FIG. 3 – Droite : Isodensité des 4 gaussiennes pour le modèle $\lambda D_k A_k D'_k$, Gauche : Classement des sondes en 4 groupes avec la règle du MAP

4 Modélisation des données avec prise en compte des connaissances biologiques

4.1 Modélisation

Afin de modéliser au mieux les données, nous ajoutons des contraintes aux modèles détaillés section 3.3. Les contraintes supplémentaires sont déduites de connaissances biologiques que nous avons sur les données. En effet, nous avons vu dans la section 2 que l'on s'attend à identifier 4 groupes différents lorsqu'on analyse des données de ChIP-chip IP/IP. Le nombre de composants du modèle de mélange est donc fixé à K=4. De plus, nous avons certaines connaissances sur les 4 groupes que l'on souhaite identifier : le groupe d'ADN non immunoprécipité et le groupe normal ont la même orientation proche de la première bissectrice. D'autre part, on suppose que le bruit est égal dans chaque groupe, ce qui revient à fixer la deuxième valeur propre de Σ_k. En effet, la première valeur propre est associée au grand axe de l'ellipse et la deuxième est associée au petit axe de l'ellipse.

Cette dernière hypothèse est utile pour contraindre la modélisation car des variances hétéroscédastiques donnent souvent des résultats très instables et ne permettent pas de retrouver les 4 groupes de la Figure 1.
Nous reprenons la paramétrisation définie section 3.2 :

$$\Sigma_k = \lambda_k D_k A_k D_k'.$$

Afin d'avoir le même bruit dans chaque groupe, on contraint la seconde valeur propre de Σ_k à être constante dans les 4 groupes. Les deux groupes qui ont la même orientation auront la même matrice D. En utilisant la décomposition des matrices de variance et sous nos contraintes, on obtient donc :

$$\begin{cases} \Sigma_k &= \lambda_k D_k A_k D_k' = D_k \Lambda_k D_k', \text{ pour } k = 1, .., 4, \text{ avec } \Lambda_k = \lambda_k A_k \\ D_1 &= D_2 = D \\ \Lambda_k &= \begin{pmatrix} u_{1k} & 0 \\ 0 & u_2 \end{pmatrix}, \text{ avec } u_{1k} > u_2, \text{ pour } k = 1, .., 4. \end{cases}$$

De manière plus générale, on peut écrire :

$$\begin{cases} \Sigma_k &= D_k' \Lambda_k D_k \ si \ k \geq 2 \\ \Sigma_k &= D' \Lambda_k D \ si \ k < 2, \end{cases}$$

où les groupes 1 et 2 correspondent aux groupes de même orientation (groupe normal et groupe d'ADN non immunoprécipité) et la matrice Λ_k est une matrice diagonale qui contient les valeurs propres de Σ_k.

L'originalité de ce modèle est de proposer la possibilité d'avoir certains composants avec une orientation fixe et d'autres composants avec une orientation libre. De plus il est possible de fixer seulement l'une des deux valeurs propres dans le choix du volume et de la forme pour un même composant du modèle. Dans le logiciel MIXMOD [4], le choix de fixer ou pas le volume, l'orientation ou la forme est obligatoirement le même pour tous les composants du modèle.

4.2 Estimation des paramètres par l'algorithme EM

Si le label de chaque donnée était observé, l'estimation des paramètres du mélange serait évidente puisque les paramètres de chaque composant $\phi(x_i; \mu_k, \Sigma_k)$ seraient estimés avec les individus de la population k. Mais les labels sont inconnus et l'estimation ne peut être fondée que sur les données observées $x_1, ..., x_n$. Il n'existe pas de formules explicites pour les estimateurs des paramètres d'un mélange, on a besoin de procédures d'estimation itératives. Le vecteur de paramètres $\Psi = (\pi_1, ..., \pi_3, \mu_1, ..., \mu_4, \Sigma_1, ..., \Sigma_4)$ est estimé à l'aide de l'algorithme EM.
Pour trouver l'estimateur des matrices de variance Σ_k, il faut maximiser l'espérance de la log-vraisemblance des données complétées en Σ_k, ce qui revient à minimiser F en D, D_k et Λ_k, où F est définie par :

$$F = \sum_{k=1}^{2} tr(D' W_k D \Lambda_k^{-1}) + \sum_{k=3}^{4} tr(D_k' W_k D_k \Lambda_k^{-1}) + \sum_{k=1}^{4} n_k \log\left\{\det(\Lambda_k)\right\},$$

où $W_k = \sum_{i=1}^{n} \tau_{ik}(x_i - \bar{x}_k)(x_i - \bar{x}_k)'$.
On remarque que seul Λ_k est présent dans les 3 termes de F. Pour D et D_k, minimiser F revient simplement à minimiser le terme où ils apparaissent. L'estimateur de D_k pour $k = 3, 4$ est le même que celui proposé par Celeux et Govaert [10] pour des composants d'orientations différentes, c'est-à-dire $\hat{D}_k$ est la matrice des vecteurs propres de W_k.

Proposition 1 *Soit* $W_k = \sum_{i=1}^{n} \tau_{ik}(x_i - \bar{x}_k)(x_i - \bar{x}_k)'$ *est une matrice de la forme* $\begin{pmatrix} w_{1k} & w_{2k} \\ w_{2k} & w_{4k} \end{pmatrix}$. *L'estimateur du maximum de vraisemblance de la matrice d'orientation* D *identique pour les deux premiers composants est de la forme* $\begin{pmatrix} \sqrt{\hat{d}} & -\sqrt{1-\hat{d}} \\ \sqrt{1-\hat{d}} & \sqrt{\hat{d}} \end{pmatrix}$, *où* $\hat{d}$ *est un réel positif défini par :*

$$\hat{d} = \begin{cases} \frac{1}{2} + \frac{\sum_{k=1}^{2}(w_{1k} - w_{4k})}{2\left\{\sqrt{(\sum_{k=1}^{2}(w_{1k} - w_{4k}))^2 + 4(\sum_{k=1}^{2}(w_{2k}))^2}\right\}} & si \ \sum_{k=1}^{2}(w_{1k} - w_{4k}) > 0 \\ \frac{1}{2} - \frac{\sum_{k=1}^{2}(w_{1k} - w_{4k})}{2\left\{\sqrt{(\sum_{k=1}^{2}(w_{1k} - w_{4k}))^2 + 4(\sum_{k=1}^{2}(w_{2k}))^2}\right\}} & sinon. \end{cases} \tag{2}$$

Idée de la preuve 1 *Minimiser* F *en* D *revient à minimiser* $f(D) = \sum_{k=1}^{2} tr(D\Lambda_k^{-1}D'W_k)$. *On peut réécrire* $f(D)$ *sous la forme suivante :*

$$f(D) = \sum_{k=1}^{2}\left(\frac{d_1' W_k d_1}{u_{1k}} + \frac{d_2' W_k d_2}{u_2}\right),$$

où d_1' *est le premier vecteur de la matrice* D *et* d_2' *le second.*
Puisque D *est une matrice orthogonale et normée, elle est de la forme* $\begin{pmatrix} \sqrt{d} & -\sqrt{1-d} \\ \sqrt{1-d} & \sqrt{d} \end{pmatrix}$. *En développant* $f(D)$ *et en dérivant par rapport à* d, *on obtient un pôlynome de degré 4 en* d *qui se résout facilement. On remarque alors que* D *ne dépend plus de* Λ. *Ce résultat analytique n'est valable qu'en dimension 2.* ■

Proposition 2 *Soit* B_k *la matrice définie par* $B_k = D_k' W_k D_k$ *de la forme* $\begin{pmatrix} b_{1k} & b_{3k} \\ b_{4k} & b_{2k} \end{pmatrix}$. *L'estimateur du maximum de vraisemblance de* Λ_k *est de la forme* $\begin{pmatrix} \hat{u}_{1k} & 0 \\ 0 & \hat{u}_2 \end{pmatrix}$, *où*

$$\begin{cases} \hat{u}_{1k} &= b_{1k}/n_k \\ \hat{u}_2 &= \sum_{k=1}^{4} b_{2k}/n \end{cases} \tag{3}$$

Idée de la preuve 2 *En développant la trace et le déterminant, on peut réécrire* F *sous la forme :*

$$F = \sum_{k=1}^{4}(b_{1k}u_{1k}^{-1} + b_{2k}u_2^{-1}) + \sum_{k=1}^{4} n_k \left\{\log(u_{1k}) + \log(u_2)\right\},$$

et minimiser F *en* Λ_k *revient à minimiser* F *en* u_{1k} *et* u_2. ■

L'estimateur de Σ_k est donc :

$$\hat{\Sigma}_k = \begin{cases} \hat{D}'_k \hat{\Lambda}_k \hat{D}_k \text{ si } k \geq 2 \\ \hat{D}' \hat{\Lambda}_k \hat{D} \text{ si } k < 2, \end{cases}$$

avec $\hat{D}$ défini par (2), $\hat{D}_k$ est la matrice des vecteurs propres de W_k et $\hat{\Lambda}_k$ défini par (3).

5 Application sur un jeu de données réel

Nous appliquons cette méthode sur les données de méthylation d'histone présentées section 2. Les données analysées concernent le chromosome 4 d'*Arabidopsis thaliana* qui est couvert par 111 699 sondes.

5.1 Initialisation de l'algorithme EM

Les résultats fournis par l'algorithme EM sont dépendants de l'initialisation. Il est important de choisir une bonne initialisation afin de ne pas tomber sur un maximum local. En pratique, on peut initialiser l'algorithme avec les résultats fournis par les différents modèles de MIXMOD [4] ou bien définir une classification initiale bien choisie. Il est souvent plus facile de définir des probabilités conditionnelles pour chaque sonde (on peut par exemple s'appuyer sur la Figure 1) que de proposer une matrice initiale Σ_k pertinente. Le critère d'arrêt choisi pour l'algorithme EM est un critère de convergence sur les paramètres avec $\varepsilon = 10^{-6}$.
Nous avons testé 11 initialisations différentes et les résultats obtenus diffèrent selon l'initialisation. Huit des 11 initialisations nous donnent le modèle auquel on s'attend biologiquement représenté schématiquement Figure 1. Mais il reste des différences : les sondes difficiles à classer qui sont au centre des 4 composants sont, selon les modèles, classées soit normales, soit appauvries, soit la moitié est classée appauvrie et l'autre moitié enrichie. Les paramètres estimés des composants ne sont alors pas les mêmes.

5.2 Critères BIC et ICL

La sélection de modèles permet de choisir le modèle minimisant le critère BIC ou le critère ICL donnés section 3.3. Le modèle minimisant à la fois le critère BIC et le critère ICL est le modèle $\lambda_k D_k A_k D'_k$ présenté Figure 2 (cf Table 1). Ce n'est pas le modèle que l'on voudrait sélectionner biologiquement. Ceci est surement dû au fait que les classes ne sont pas des gaussiennes en réalité.

5.3 Estimation des paramètres

Nous présentons les résultats du modèle initialisé avec des probabilités conditionnelles. Les paramètres du mélange estimés par l'algorithme EM sont donnés dans la Table 2. Les proportions de chacun des groupes correspondent à celles attendues par les biologistes. En effet, on sait que la méthylation de cette histone n'est présente qu'en faible proportion dans le génome. Or nous trouvons environ 39% des sondes dans le groupe non immunoprécipité. Nous savons aussi que la différence de méthylation est majoritairement appauvrie chez

	Modèle $\lambda_k D_k A_k D_k'$	Modèle $\lambda D_k A_k D_k'$	Modèle 1	Modèle 2
nb de paramètres	23	20	18	18
BIC	578 171	637 770	606 643	613 470
ICL	607 488	690 126	639 101	640 582

TAB. 1 – Critères BIC et ICL selon les modèles. Modèle 1 correspond à notre modèle initialisé avec des probabilités conditionnelles, Modèle 2 correspond à notre modèle initialisé avec des paramètres bien choisis

le mutant et très rarement enrichie. Le groupe appauvri regroupe 22% des sondes et le groupe enrichi en regroupe seulement 13%.
D'autre part, la matrice d'orientation D estimée pour les groupes 1 et 2 est très proche de la matrice d'orientation attendue pour une direction sur la première bissectrice.

	Groupe 1	Groupe 2	Groupe 3	Groupe 4
$\hat{\pi}$	0.39	0.26	0.13	0.22
$\hat{\mu}$	[7.56 ;7.54]	[12.19 ;12.04]	[8.07 ;9.17]	[9.10 ;7.95]
$\hat{D}$	$\begin{pmatrix} 0.71 & -0.70 \\ 0.70 & 0.71 \end{pmatrix}$	$\begin{pmatrix} 0.71 & -0.70 \\ 0.70 & 0.71 \end{pmatrix}$	$\begin{pmatrix} 0.32 & -0.94 \\ 0.94 & 0.32 \end{pmatrix}$	$\begin{pmatrix} -0.96 & 0.26 \\ -0.26 & -0.96 \end{pmatrix}$
$\hat{\Lambda}$	$\begin{pmatrix} 0.11 & 0 \\ 0 & 0.14 \end{pmatrix}$	$\begin{pmatrix} 9.42 & 0 \\ 0 & 0.14 \end{pmatrix}$	$\begin{pmatrix} 1.41 & 0 \\ 0 & 0.14 \end{pmatrix}$	$\begin{pmatrix} 1.29 & 0 \\ 0 & 0.14 \end{pmatrix}$
$\hat{\Sigma}$	$\begin{pmatrix} 0.12 & -0.01 \\ -0.01 & 0.12 \end{pmatrix}$	$\begin{pmatrix} 4.8 & 4.64 \\ 4.64 & 4.75 \end{pmatrix}$	$\begin{pmatrix} 0.27 & 0.39 \\ 0.39 & 1.28 \end{pmatrix}$	$\begin{pmatrix} 1.21 & 0.29 \\ 0.29 & 0.22 \end{pmatrix}$

TAB. 2 – Estimation des paramètres. Les groupes 1 et 2 correspondent aux groupes normaux (le groupe 1 est le groupe non-immunoprécipité), le groupe 3 correspond au groupe enrichi et le groupe 4 correspond au groupe appauvri.

On obtient quatre groupes en classant chaque sonde dans le groupe pour laquelle la probabilité conditionnelle est la plus grande (cf Figure 4).
D'un point de vue biologique, le plus important est dans un premier temps de distinguer les sondes enrichies ou appauvries (c'est-à-dire là où le taux de méthylation est différent entre le sauvage et le mutant). On peut donc considérer les groupes de même orientation (groupes 1 et 2) comme un seul groupe qui correspond à un taux de méthylation identique dans les deux échantillons (groupe normal). On veut donc classer les sondes en trois groupes : normal, appauvri ou enrichi. Pour cela, on somme les probabilités conditionnelles des groupes 1 et 2. Une autre possibilité est de classer en deux groupes seulement, un groupe qui correspond à une méthylation identique dans les deux échantillons, et l'autre qui correspond à un taux de méthylation différent entre les deux échantillons. Pour cela,

on somme les probabilités conditionnelles des groupes 1 et 2 et celles des groupes 3 et 4. Lorsque l'on classe en 4 groupes, on trouve bien les 4 groupes comme attendus sur la figure 1, mais il est probable que les sondes aux frontières de deux classes aient des probabilités conditionnelles très proches pour les deux classes et soient donc mal classées. Comme nous préférons ne pas avoir d'information sur une sonde plutôt que d'avoir une information fausse, nous fixons un seuil de classification à 0.7, ce qui délimite une marge de non classement autour de chacun des groupes (cf Figure 5). Avec un seuil à 0.7, seulement 12.5% des sondes ne sont pas classées. On peut bien sûr faire de même avec les classements en 2 ou 3 groupes, le nombre de sondes non classées est alors plus faible (11.9% pour un classement en 3 groupes et 9.3% pour un classement en 2 groupes).

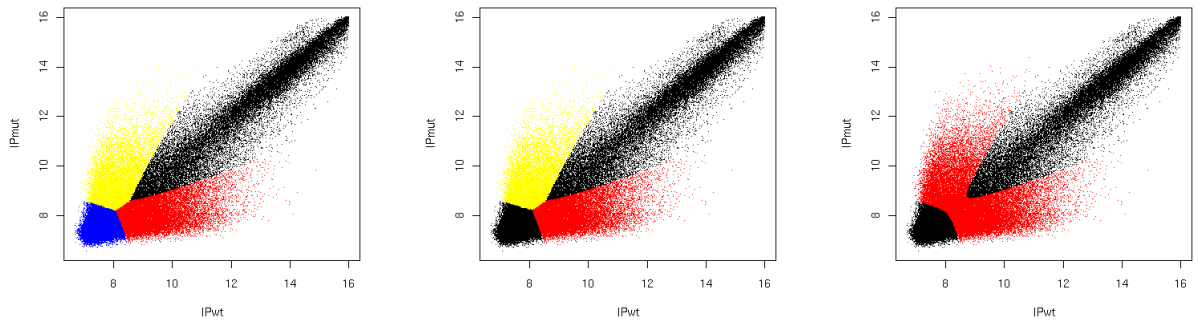

FIG. 4 – Classement des sondes en 4 groupes (gauche), 3 groupes (centre), 2 groupes (droite).

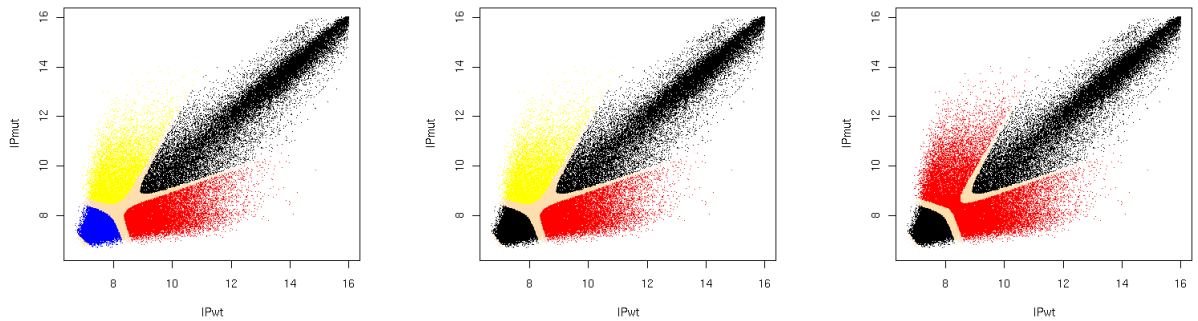

FIG. 5 – Classement des sondes en 4 groupes (gauche) avec un seuil de classification à 0.7 (zone en gris), 3 groupes (centre), 2 groupes (droite).

5.4 Interprétations biologiques

Nous avons ensuite comparé les résultats à l'annotation connue d'*Arabidopsis thaliana* à l'aide du logiciel $SignalMap^{TM}$ fourni par NimbleGen (cf Figure 6). Bien que notre modèle ne prenne pas en compte la structure spatiale des sondes le long du chromosome, les sondes déclarées normales, enrichies ou appauvries chez le mutant sont regroupées sous forme de plage. On s'attend évidemment à ce que des sondes contigües aient le même comportement. D'autre part, la marque H3K9me2 étudiée est une marque hétérochromatinienne présente sur environ 15% du génome. La plupart des régions couvertes par H3K9me2 sont contigües et couvrent plusieurs mégabases dans les régions péricentromériques ou dans l'hétérochromatine interstitielle comme le knob du chromosome 4, mais il existe aussi des régions plus petites (ilôts d'hétérochromatine) situées dans l'euchromatine et qui couvrent majoritairement des éléments transposables (Bernatavichute *et al.* [3]). Nous savons aussi qu'il y a peu de différences entre le sauvage et le mutant pour le taux de méthylation d'H3K9me2. Nos résultats corroborent parfaitement ces connaissances. En effet, on observe une majorité de sondes déclarées non méthylées le long du chromosome 4, mais dans la région péricentromérique (entre les positions 2 800 000 et 5 000 000) et autour du knob (entre les positions 1 600 000 et 2 300 000), on remarque une majorité de sondes du groupe normal et des larges plages de sondes du groupe appauvri ou enrichi. On détecte aussi des plages de sondes, plus petites, appartenant au groupe enrichi ou appauvri situées dans l'euchromatine et qui couvrent des éléments transposables (cf Figure 6). Environ 10% des sondes du génome couvrent des éléments transposables, et on trouve 26% des sondes du groupe normal couvrant un élément transposable. Un test du χ^2 montre une différence significative. Il y a clairement un biais et on peut donc dire que la marque H3K9me2 est majoritairement présente sur les éléments transposables.

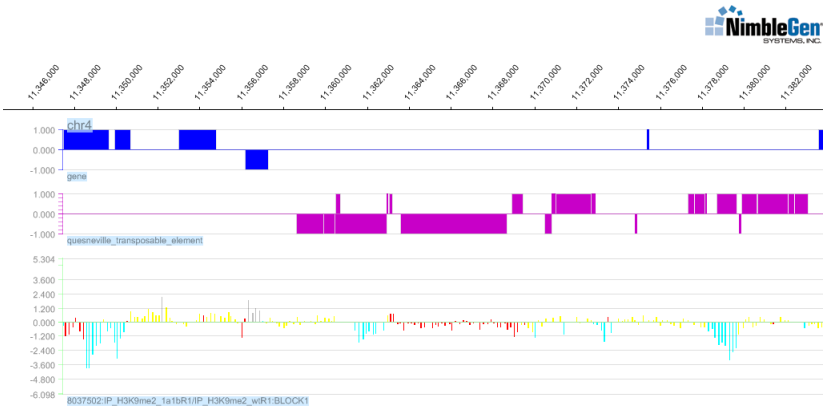

FIG. 6 – Comparaison à l'annotation. En bleu les gènes (1ère ligne), en violet les éléments transposables (2ème ligne). En rouge, les sondes où la méthylation est identique entre le sauvage et le mutant. En noir les sondes déclarées enrichies, en bleu les appauvries, en jaunes les non-méthylées (3ème ligne).

6 Conclusion

Nous proposons une méthode fondée sur un mélange de gaussiennes bidimensionnelles contraintes pour l'analyse de données de ChIP-chip IP/IP. La connaissance biologique des données est prise en compte. Les paramètres sont estimés par l'algorithme EM. Cette méthode donne des résultats convaincants pour l'analyse d'un jeu de données réel concernant la méthylation d'une histone. Nous souhaitons aussi analyser d'autres types de données où il n'y aurait que 3 groupes à définir (pas d'appauvri, pas d'enrichi ou pas de non-immunoprécipité). Bien que notre modèle ne prenne pas en compte la structure spatiale des sondes le long du chromosome, les sondes déclarées normales, enrichies ou appauvries sont regroupées sous forme de plage. Une amélioration naturelle consiste à prendre en compte la structure spatiale des sondes en utilisant un modèle de type HMM. D'autre part, on peut aussi rajouter des contraintes de symétrie entre les groupes appauvri et enrichi.

Références

[1] Amaratunga, D. and Cabrera, J. : Exploration and Analysis of DNA Microarray and Protein Array Data. *Wiley Series in Probability and Statistics* (2004).

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

[3] Bernatavichute, Y.V., Zhang, X., Cokus, S., Pellegrini, M. and Jacobsen, S.E. : Genome-Wide Association of Histone H3 Lysine Nine Methylation with CHG DNA Methylation in Arabidopsis thaliana. *PLoS ONE* **3(9) :e3156** (2008).

[4] Biernacki, C., Celeux, G., Echenim, A., Govaert, G. and Langrognet, F. : Le logiciel MIXMOD d'analyse de mélange pour la classification et l'analyse discriminante. *La Revue de Modulad* **35** (2007) 25-44.

[5] Biernacki, C., Celeux, G. and Govaert, G. : Assessing a mixture model for clustering with the integrated completed likelihood. *Pattern Analysis and Machine Intelligence* **22(7)** (2000) 719-725.

[6] Boulicaut, J.F. and Gandrillon O. : Informatique pour l'analyse du transcriptome. *Lavoisier* (2004).

[7] Buck, M.J. and Lieb, J.D. : Chip-chip : considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics* **83** (2004) 349-360.

[8] Cappé, O., Moulines, E. and Rydén, T. : Inference in hidden Markov models. *Springer Series in Statistics, NY : Springer* (2005).

[9] Cawley, S. *et al.* : Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116** (2004) 499-509.

[10] Celeux, G. and Govaert, G. : Gaussian Parsimonious Clustering Models. *Pattern Recognition* **28** (1995) 781-793.

[11] Dempster, A.P., Laird, N.M. and Rubin, D.B. : Maximum likelihood from incomplete data via the EM algorihtm (with discussion). *J. R. Statis. Soc. B* **39** (1977) 1-38.

[12] Ephraim, Y. and Merhav, N. : Hidden Markov processes. *IEEE Transactions on Information Theory* **48(6)** (2002) 1518-1569.

[13] Hupé, P., Stransky, N., Thiery, JP., Radvanyi, F. and Barillot, E. : Analysis of array CGH data : from signal ratio to gain and loss of DNA regions. *Bioinformatics* **20(18)** (2004) 3413-22.

[14] Keles, S. : Mixture modeling for genome-wide localization of transcription factors. *Biometrics* **63** (2007) 10-21.

[15] Martin-Magniette, M-L, Mary-Huard, T., Berard, C. and Robin S. : ChIPmix : mixture model of regressions for two-color ChIP-chip analysis. *Bioinformatics* **24 :i181-i186** (2008).

[16] Olshen, AB., Venkatraman, ES., Lucito, R. and Wigler, M. : Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5(4)** (2004) 557-72.

[17] Picard, F., Robin, S., Lavielle, M., Vaisse, C. and Daudin, JJ. : A statistical approach for array CGH data analysis. *BMC Bioinformatics* **6 :27** (2005).

[18] Turck, F., Roudier, F., Farrona, S., Martin-Magniette, M-L. *et al.* : Arabidopsis TFL2/LHP1 Specifically Associates with Genes Marked by Trimethylation of Histone H3 Lysine 27. *PLoS Genet.v* **3 :6** (2007).

()