Mélange de gaussiennes bidimensionnelles pour l'analyse de données de ChIP-chip IP/IP

Caroline Bérard, Marie-Laure Martin-Magniette, François Roudier, Stéphane Robin

Mélange de gaussiennes bidimensionnelles pour l'analyse de données de ChIP-chip IP/IP.

Caroline Bérard[1], Marie-Laure Martin-Magniette[1,2], François Roudier[3] & 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, Ecole Normale Supérieure, 46 rue d'Ulm, PARIS Cedex 05.*

Résumé : Le ChIP-chip (immunoprécipitation de la chromatine sur puce) est une technique utilisée pour étudier les interactions entre protéines et ADN. Elle permet notamment d'étudier la différence entre deux échantillons d'ADN immunoprécipité (issus d'un sauvage et d'un mutant par exemple). Biologiquement, on s'attend alors à distinguer quatre groupes différents : un groupe d'ADN non immunoprécipité (c'est-à-dire d'intensité faible), un groupe d'ADN immunoprécipité identiquement dans les deux échantillons, et puis deux groupes dans lesquels l'ADN est immunoprécipité différemment. Nous proposons de modéliser ces données par un mélange de gaussiennes bidimensionnelles à quatre composantes. Des contraintes sur les matrices de variance-covariance sont posées afin d'intégrer des connaissances biologiques. Les paramètres sont estimés par l'algorithme EM. Nous appliquons cette méthode sur des données issues de la technologie NimbleGen afin d'étudier la différence de méthylation d'une histone chez la plante modèle *Arabidopsis thaliana* entre l'écotype sauvage et un mutant.

Abstract : ChIP-chip (Chromatin immunoprecipitation on chip) is a well-established procedure to investigate proteins associated with DNA. ChIP-chip enables to study differences between two immunoprecipitated DNA samples (a wildtype and a mutant for example). From a biological point of view, we expect to distinguish four different groups: a group of non-immunoprecipitated DNA (with low intensity), a group of immunoprecipitated DNA in both samples, and then two groups in which DNA is immunoprecipitated differently (immunoprecipitated strongly in a sample and low in the other). We propose to model these data with a mixture of two-dimensional gaussians with four components. We incorporate biological knowledge by translating them in 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 model plant *Arabidopsis thaliana* and a mutant.

Mots-clés : Biologie-Génome, Choix de Modèle.

Le ChIP-chip (immunoprécipitation de la chromatine sur puce) est une technique utilisée pour étudier les interactions entre protéines et ADN. Habituellement dans une expérience de ChIP-chip, les deux échantillons co-hybridés sont les fragments d'ADN liés à la protéine d'intérêt (IP) et l'ADN génomique total (INPUT). Le but est alors d'identifier l'ADN lié à la protéine d'intérêt, c'est-à-dire les sondes qui ont un signal IP plus fort que le signal INPUT, appelées alors sondes enrichies.
Buck et Lieb (2004) 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.* 2004, Keles 2007), 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 2004, Turck *et al.* 2007, Martin-Magniette *et al.* 2008). Différentes méthodes statistiques ont été proposées pour distinguer ces deux populations: toutes sont fondées sur la distribution du log-ratio (Buck et Lieb 2004, Turck *et al.* 2007), exceptée la méthode proposée par Martin-Magniette *et al.* (2008) 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 (issus d'un sauvage et d'un mutant par exemple), sans hybrider sur la puce l'ADN génomique total (INPUT). On s'attend alors à distinguer quatre groupes différents (cf graphe 1) : un groupe d'ADN non-immunoprécipité (c'est-à-dire d'intensité faible), un groupe d'ADN immunoprécipité identiquement dans les deux échantillons (groupe normal), et puis deux groupes dans lesquels l'ADN est immunoprécipité différemment (immunoprécipité fortement dans un échantillon et faiblement dans l'autre).

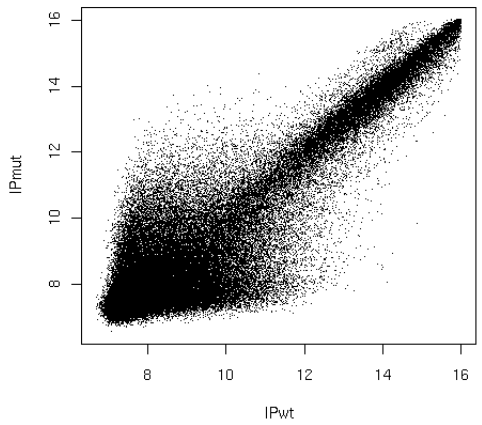

Figure 1: Données de méthylation d'ADN entre un sauvage et un mutant

À notre connaissance il n'existe pas de méthode pour analyser ce type de données dans la littérature. Les méthodes de segmentation initialement développées pour l'analyse des données CGH (Hupé 2004, Olshen 2004, Picard 2005) pourraient être utilisées, mais seuls 3 groupes pourraient être caractérisés car les régions génomiques non-immunoprécipitées et les régions immunoprécipitéés dans les deux échantillons seraient indistinguables. De plus ces méthodes sont assez coûteuses en temps de calcul dans le cas où il y a environ 200 000 sondes par chromosome.
Dans ce travail, nous proposons de modéliser les données par un mélange de gaussiennes bidimensionnelles à quatre composantes. Soit $X_i = (x_{1i}, x_{2i})$ le signal log-IP de chaque échantillon pour la sonde i respectivement, la densité du couple s'écrit :

$$f(X_i) = \sum_{k=1}^{4} \pi_k \phi(X_i|\mu_k, \Sigma_k),$$

où π_k est la proportion de la $k$ème composante du mélange ($0 < \pi_k < 1$, $\forall\ k = 1, ..., 4$ et $\sum_{k=1}^{4} \pi_k = 1$) et $\phi(.|\mu_k, \Sigma_k)$ est la densité d'une distribution gaussienne bidimensionnelle de paramètres (μ_k, Σ_k), où μ_k est la moyenne et Σ_k est la matrice de variance-covariance. ϕ est définie par :

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

Le vecteur des paramètres du mélange est $\theta = (\pi_1, ..., \pi_3, \mu_1, ..., \mu_4, \Sigma_1, ..., \Sigma_4)$.

Afin d'intégrer la connaissance biologique, nous ajoutons des contraintes au modèle. En effet, 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. De plus, on pose une contrainte de variance : on suppose que l'on a le même bruit dans chaque groupe (variances égales). Nous reprenons donc une paramétrisation proposée par Banfield et Raftery (1993) qui ont considéré la décomposition spectrale des matrices de variance des classes :

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

où λ_k représente le volume ($\lambda_k = \det(\Sigma_k)^{1/2}$), D_k représente l'orientation et A_k la forme. 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. Cette paramétrisation permet de proposer de nombreux modèles de classification (Celeux et Govaert (1995)), implémentés dans le logiciel MIXMOD (Biernacki *et al.* (2007)).
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 :

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

L'expression de l'espérance de la log-vraisemblance de notre modèle est :

$$\begin{aligned} Q &= \sum_{k=1}^{4} n_k \log(\pi_k) - \frac{1}{2} pn \log(2\pi) + \frac{1}{2} \sum_{k=1}^{4} n_k \log\left[\det\{(D_k \Lambda_k D_k')^{-1}\}\right] \\ &\quad - \frac{1}{2} \sum_{k=1}^{4} tr\{(D_k \Lambda_k D_k')^{-1} W_k\} - \frac{1}{2} \sum_{k=1}^{4} n_k (\bar{x_k} - \mu_k)^T (D_k \Lambda_k D_k')^{-1} (\bar{x_k} - \mu_k). \end{aligned}$$

Les paramètres $(\pi_1, ..., \pi_3, \mu_1, ..., \mu_4, \Sigma_1, ..., \Sigma_4)$ sont estimés à l'aide d'un algorithme E-M. Dans l'étape M, trouver l'estimateur de Σ_k revient à trouver les estimateurs de D_k et Λ_k qui sont calculés en minimisant

$$F = \sum_{k=1}^{4} tr(D_k' W_k D_k \Lambda_k^{-1}) + \sum_{k=1}^{4} n_k \log\{\det(\Lambda_k)\}.$$

On veut ensuite attribuer à chaque sonde le groupe pour lequel elle a la plus forte probabilité d'appartenance. Pour cela, on calcule les probabilités conditionnelles que la sonde i appartienne au groupe k sachant l'ensemble des observations.

$$\tau_{ik} = \frac{\hat{\pi}_k \phi(X_i | \hat{\mu}_k \hat{\Sigma}_k)}{\sum_{l=1}^{4} \hat{\pi}_l \phi(X_i | \hat{\mu}_l \hat{\Sigma}_l)}$$

On utilise la règle du *Maximum A Posteriori* et chaque sonde est finalement classée dans le groupe pour laquelle la probabilité conditionnelle est la plus grande.
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 l'immunoprécipitation est différente entre le sauvage et le mutant). Pour cela, on somme les probabilités conditionnelles des deux groupes de même orientation (groupes 1 et 2) pour constituer un seul groupe : le groupe normal (l'immunoprécipitation est la même dans les deux échantillons). Puis on classe les sondes en trois groupes (normal, appauvri ou enrichi) selon la règle du MAP.
Nous appliquons cette méthode sur des données issues de la technologie NimbleGen afin d'étudier la différence de méthylation d'une histone entre la plante modèle *Arabidopsis thaliana* et un mutant.

Bibliographie

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

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

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

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

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

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

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

[8] Picard F. *et al.* (2005). A statistical approach for array CGH data analysis. *BMC Bioinformatics* 6:27.

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

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

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