Nouvelle méthode statistique pour l'analyse de données de ChIP-chip

Florian Salipante, Christelle Reynes, Laurent Journot, Robert Sabatier

Nouvelle méthode statistique pour l'analyse de données de ChIP-chip

Florian Salipante[1,3], Christelle Reynes[2], Laurent Journot[3] & Robert Sabatier[1]

[1] *Laboratoire de Physique Industrielle et Traitement de l'Information, EA 2415, Faculté de Pharmacie 15 Avenue Charles Flahault BP 14491, 34093 MONTPELLIER Cedex 5, France*

[2] *MTi - Unite Inserm - Paris 7 Diderot U973, Bat Lamarck 5ème étage, 35, rue Hélène Brion, 75205 Paris Cedex 13, FRANCE*

[3] *Institut de Génomique Fonctionnelle du CNRS, 141, rue de la Cardonille, 34094 MONTPELLIER Cedex 5, France*

Résumé

La méthode de Chromatin ImmunoPrecipitation on chip (ChIP on chip ou ChIP-chip) a pour but de détecter les sites de fixation des protéines (généralement des facteurs de transcription) sur la molécule d'ADN. L'analyse statistique des données consiste a rechercher des régions de pics significatifs synonymes de sites de fixation. La méthode que nous avons élaborée est issue de la théorie des valeurs extrêmes et particulièrement de la méthode POT (Peaks Over Threshold). Cette méthode consiste à modéliser les données de queues de distribution, en ne retenant que les valeurs dépassant un certain seuil μ, elle a la particularité de modéliser d'une part les intensités de dépassement de seuil, mais aussi les positions d'occurrences de ces dépassements de seuil. Cette méthode va nous permettre de déterminer un seuil au delà duquel les pics pourront être considérés comme significatifs.
Mots-clés : ChIP-chip, Modèle POT, puce à ADN, détection de pics, loi de Pareto généralisée, loi binomiale négative

Abstract

Chromatine ImmunoPrecipitation on chip (ChIP on chip or ChIP-chip) technology is used to detect protein (transcription factors generally) binding site on DNA. The statistical analysis consists in looking for significant peaks regions, in order to find binding sites. The method which we elaborated come from the extreme value theory and especially from the POT method (Peaks Over Threshold). This method consists in distribution tails modelling, by only retaining the values exceeding a given threshold μ, it has the peculiarity that it modelize on one hand the intensities of excesses over threshold and on the other hand the occurrences of these excesses over threshold. This method allow us to determine a threshold beyond which peaks can be considered as significant.
Keywords : ChIP-chip, POT, microarray, peak detection, Generalized Pareto distribution, Negative Binomial distribution

Introduction

La méthode ChIP-chip utilisée en génomique est relativement récente, elle a pour but de détecter les sites de fixation des protéines sur la molécule d'ADN. Certaines de ces protéines, les facteurs de transcription, ont un rôle prépondérant dans le contrôle de la transcription de nos gènes. Ces expériences peuvent être réalisées à l'échelle d'un chromosome ou du génome ce qui implique des données de très grande dimension (avec des puces Affimetrix, 90 millions de sondes espacées tout les 35 paires de bases environ permettent l'étude du génome humain dans son intégralité). L'analyse statistique de ces données consiste à rechercher, à l'échelle du génome, les régions comportant des pics significatifs (synonymes de sites de fixation) en tenant compte, d'une part de la dépendance locale des données inhérente à ce type d'expérience et d'autre part des nombreux artefacts présents liés au caractère "bruité" de ce type de données. Une première approche naturelle consiste à lisser les données par moyenne mobiles par exemple, comme le font la majorité des méthodes statistiques appliquées à ce type de données. La méthode que nous avons élaborée est issue du constat qu'à l'échelle du génome, les sites de fixation sont des évènements rares. Ce qui nous a conduit à utiliser une méthode issue de la théorie des valeurs extrêmes, la méthode POT pour Peaks Over Threshold (Beirlant *et al.*, 2004). Cette méthode consiste à modéliser les données de queues de distribution, en ne retenant que les valeurs dépassant un certain seuil μ, elle a pour particularité le fait qu'elle modélise d'une part les intensités de dépassement de seuil, mais aussi les positions des occurrences de ces dépassements de seuil. Cette méthode originale va nous permettre de déterminer un seuil au delà duquel les pics pourront être considérés comme significatifs, mais aussi de tenir compte du caractère de la zone génomique étudiée. Par exemple, il est possible que les facteurs de transcription soient plus présents dans des zones où il y a une accumulation de nucléotides C et G (CpG islands).

1 Les données

Le principe de fonctionnement de la technique ChIP-chip se décompose en deux parties, la partie $ChIP$ pour Chromatine ImmunoPrecipitation et la partie *chip* qui correspond à une étude sur puce à ADN. La partie $ChIP$ consiste dans un premier temps à fixer les protéines à l'ADN *in vivo* à l'aide de formaldéhyde. Vient en suite l'étape de "sonication" où l'ADN est fragmenté aléatoirement, puis à l'aide d'anticorps spécifiques intervient le processus d'ImmunoPrecipitation qui permet de retenir les fragments porteurs de la protéine étudiée. L'ADN et la protéine sont alors séparés. Dans la partie *chip*, on récupère des fragments d'ADN provenant de deux expériences, une où les fragments d'ADN ont subis l'étape d'ImmunoPrecipitation et l'autre non, chacune des expériences étant marquée par un fluorochrome différent. Les fragments sont alors hybridés sur une puce à ADN et les niveaux de fluorescence des sondes vont permettre de déterminer si le

fragment a été enrichi (par rapport à la protéine étudiée), c'est-à-dire s'il correspond à un site de fixation. Bien que l'on utilise une puce à ADN pour la confection de ces données, l'analyse statistique ne sera pas la même que pour les puces à ADN expression. En effet, le design en "tuilage" (proximité ou même chevauchement des sondes placées sur la puce) utilisé pour les ChIP-chips entraîne une dépendance locale des données, ce qui n'est pas le cas pour les puces à ADN expression où chaque sonde comporte des fragments éloignés. En effet, il serait fort surprenant que deux sondes voisines aient des valeurs complètement différentes alors même que les fragments qui les composent sont très proches voire se chevauchent sur le génome. Ainsi, toute position ayant un fort niveau de fluorescence, mais ayant des valeurs voisines avec un faible niveau de fluorescence ne devra pas être considérée par l'analyse statistique comme un site de fixation mais comme un artefact. Les données $x_1, ..., x_n$ se présentent donc sous forme d'un vecteur de taille n, égal au nombre de sondes (cela peut varier de quelques milliers pour l'étude d'un chromosome, à plusieurs millions pour le génome humain tout entier). L'objectif va être de détecter des pics significatifs, synonymes de sites de fixation.

2 Matériels et méthodes

2.1 Etat de l'art

Il existe plusieurs manières d'aborder le problème de l'analyse des données de ChIP-chip. La plus naturelle consiste à lisser le signal de manière à retirer les artefacts, plusieurs méthodes utilisent cette technique comme Buck & Lieb (2003), Glynn *et al.* (2004) et Buck *et al.* (2005), le problème de ces méthodes est que le choix de la taille de la fenêtre mobile de lissage et du pas sont déterminés de manière arbitraire, de plus, les zones identifiées comme des sites de fixation sont déterminées à l'aide du calcul de p-values obtenues avec un test ayant pour hypothèse nulle que le log ratio des deux populations suit une loi normale centrée en zéro (suivi d'une correction pour tests multiples de type Bonferroni). Ces techniques apparaissent comme beaucoup trop rudimentaires pour l'analyse de données aussi complexes que le sont les ChIP-chips.
Dans Cawley *et al.* (2004), les auteurs réalisent un test de Wilcoxon de comparaison entre la distribution de la population ayant subie l'ImmunoPrecipitation contre celle ne l'ayant pas subie, dans une fenêtre mobile le long du génome. Là encore, la définition de la taille de la fenêtre ainsi que la méthode statistique employée semblent être déterminés de manière assez arbitraire.
Li *et al.* (2005) et Ji & Wong (2005) utilisent des chaînes de Markov cachées pour la recherche des sites de fixation. Les deux états cachés sont l'état enrichi (sous entendu ayant subi l'ImmunoPrecipitation) et l'état non enrichi. La séquence la plus probable est calculée à l'aide des algorithmes habituels "forward" et "backward". L'inconvénient majeur de cette méthode réside dans le fait qu'il est nécessaire d'avoir un *a priori* sur le nombre de sites de fixation potentiels pour le calcul des probabilités initiales et de

transitions, ce qui est rarement le cas.
Enfin, Zheng *et al.* (2007) modélisent les pics par des triangles obtenus par régression linéaire multiple. Une valeur moyenne pour chaque triangle modélisé est calculée, si les valeurs d'un triangle donné ne proviennent pas d'une région génomique où il y a un site de fixation, alors elles doivent être assimilées a du bruit modélisable par un processus stationnaire. La valeur moyenne de chaque triangle sera donc comparée à une loi normale centrée et de variance appropriée à l'échantillon. La p-value obtenue permettra après définition d'un seuil, d'écarter les faux sites de fixation, cependant la détermination du seuil n'est pas explicitement décrite dans l'article.

2.2 Notre méthode

La méthode que nous avons mise au point est issue du constat qu'à l'échelle du génome, quelle que soit la protéine étudiée, les sites de fixation de cette protéine peuvent être considérés comme des évènements rares. Cela nous conduit naturellement vers la théorie des valeurs extrêmes et plus précisément vers le modèle POT (Peaks Over Threshold). Cette méthode à l'origine utilisée en météorologie ou encore en finance a pour particularité de modéliser les queues de distribution, en ne conservant que les données dépassant un seuil μ. Elle modélise d'une part l'intensité de ces dépassements de seuil mais aussi leur position d'occurrence. Cependant, cette méthode n'est pas applicable aux données en tant que telle, un certain nombre de modifications ainsi qu'un prétraitement des données va être nécessaire afin qu'elles puissent rentrer dans le cadre d'application.

2.2.1 Filtage des données

Le filtrage de nos données est réalisé de manière à enlever un maximum d'artefacts qui pourraient biaiser la modélisation future. Pour ce faire, on utilise le *super smoother* de Friedman (Friedman, 1984), pour effectuer un lissage des données. Si a une position donnée i la valeur d'origine associée x_i dépasse sa valeur lissée x'_i et que dans son voisinage proche ($i-2 : i+2$ par exemple) aucune valeur ne dépasse le signal lissé, alors on considérera que la valeur x_i est un artefact et on lui affectera comme valeur la moyenne des valeurs de son voisinage ($\frac{1}{4}\sum_{j=-2;j\neq 0}^{2} x_{i+j}$). Cette étape n'est qu'un premier "filtrage" des données, elle a pour but de retirer les artefacts les plus évidents, en prenant bien soin de ne pas être trop restrictif et d'éviter de retirer des valeurs qui pourraient réellement être des sites de fixation.

2.2.2 Extraction des pics

L'utilisation du modèle POT requiert des données indépendantes, mais, comme on l'a vu précédemment, ce n'est pas le cas pour les ChIP-chips à cause du design en "tuilage" des puces à ADN. Nous allons donc utiliser un algorithme d'extraction des pics. Au terme

de l'algorithme, on a m pics décrits par leurs maxima, les positions des maxima et les positions des bornes. On peut espérer que les données sélectionnées seront suffisamment éloignées sur le génome pour ne plus être dépendantes, et on a pu vérifier sur plusieurs jeux de données que ces valeurs n'étaient pas corrélées, ceci n'étant pas une condition suffisante mais néanmoins nécessaire. Nous allons donc délimiter des zones ne contenant qu'un seul et unique pic, significatif ou non. Pour cela, on utilise à nouveau le *super smoother* de Friedman pour lisser les données et on procède comme suit :

- Au départ, on a une seule zone qui s'étend sur l'ensemble des données,
- on cherche la valeur maximale de cette zone ($\max_{i\in\{1,...,n\}} x_i$),
- on détermine les bornes, à l'aide du signal lissé (x'), pour la borne de gauche, on se déplace vers la gauche et on calcule pour chaque position i la dérivée numérique ($\frac{x'_{i+1}-x'_{i-1}}{(i+1)-(i-1)}$), quand celle-ci redevient négative, on atteint nouvelle borne, la procédure est analogue pour la borne de droite,
- on obtient deux zones, de la 1^{ere} position à la borne de gauche, et de la borne de droite à la dernière position et on réitère le processus jusqu'à qu'il n'y ai plus de zones où que celles ci soient trop petites pour contenir un site de fixation.

Les données utilisées dans le modèle POT seront donc les intensités des maximas obtenus $(y_1, ..., y_m)$ ainsi que leurs positions $(p_1, p_2, ..., p_m)$, chaque donnée est donc un couple.

2.2.3 Le modèle POT

En fonction du seuil μ, parmis les m maxima, seuls l d'entres eux dépasseront le seuil. Avec la méthode POT, on modélise :

- Le processus $(Z_i)_{i=1:l}$ par une loi de pareto généralisée de vraisemblance :

$$L_l^1 = \frac{1}{\sigma^l}\prod_{i=1}^{l}(1+\frac{\xi(z_i-\mu)}{\sigma})^{(-\frac{1}{\xi}-1)}$$

  , où σ, ξ et μ sont respectivement des paramètres d'échelle, de forme et de position, les z_i étant les intensités des maxima dépassant le seuil μ,

- N (la variable occurence des dépassements de seuil) par une loi binomiale négative non homogène de vraisemblance (habituellement une loi de poisson non homogène est utilisée, mais la loi binomiale négative est une bonne alternative quand celle-ci ne convient pas) :

$$L_l^2 = \frac{1}{(1+\frac{\Lambda(P)}{R(P)})^{R(P)}}\prod_{i=1}^{l}\Lambda(P_i)(\frac{R(P_i)}{R(P_i)+\Lambda(P_i)})^{R(P_i)+1}$$

Où

$$\Lambda(S) = \int_{p_i}^{p_j} \lambda(p)dp, \quad R(S) = \int_{p_i}^{p_j} r(p)dp \quad si \quad S = [p_i, p_j]$$

$P_1, ..., P_2, ..., P_l$ sont les positions où ont lieu les l évènements de dépassement de seuil, P étant un regroupement des $m - l$ positions restantes. Enfin $\lambda(p)$ et $r(p)$ sont les paramètres de la loi binomiale négative non homogène dépendant de la position p.

La valeur du seuil μ est déterminée de manière à maximiser la vraisemblance du modèle POT. Etant donné que les processus y_i et p_i n'ont aucune raison d'être dépendants, la log vraisemblance du modèle POT peut s'écrire comme suit : $l_m^P = l_m^1 + l_m^2$
l_m^1 étant la log vraisemblance de la loi GPD et l_m^2 la log vraisemblance de la loi binomiale négative non homogène.
La dernière étape consiste à déterminer les régions significatives. Pour cela, on revient aux données initiales $x_1, x_2, ..., x_n$, parmi les m positions où il y a un maximum, l dépassent le seuil. Considérons x_i comme une de ces l valeurs, la zone de x_i définie lors de l'étape d'extraction des pics sera considérée par cette méthode comme étant un site de fixation si : $(\sum_{j=-2;j\neq 0}^{2} \mathbf{1}_{\{x_{i+j}>\mu\}}) \geq 1$ (où $\mathbf{1}$ est la fonction indicatrice), c'est à dire si au moins une autre valeur dans son voisinage proche dépasse aussi le seuil μ.
A l'oral nous présenterons des résultats obtenues sur des données réelles, ainsi que sur des données de la littérature.

Conclusion

La dérégulation de l'expression des gènes est la cause de nombreuses pathologies. La technologie ChIP-chip représente une avancée considérable pour l'identification de complexes de facteurs de transcription, ce qui permettra une meilleure compréhension du système de régulation sous-jacent aux pathologies. Cependant, ces données au caractère "gigantesque" et "bruité" nécessitent un traitement statistique approprié pour être efficaces. Nous avons mis au point une méthode inspirée du modèle POT, qui considère les sites de fixation d'un facteur de transcription donné, comme des événements rares à l'échelle du génome. Elle permet l'identification de ces sites par la recherche d'un seuil au delà duquel un pic pourra être considéré comme significatif.
Plus récemment, une nouvelle technologie est apparue sous le nom de ChIP-Sequencing (ChIP-Seq), c'est une méthode alternative aux ChIP-chip pour l'étude de l'interaction protéine-ADN lors de la phase de transcription. Les données issues de ce nouveau procédé semblent être de nature sensiblement analogue à celle obtenues avec des ChIP-chip, la prochaine étape sera donc de transposer notre méthode à l'analyse de données ChIP-seq, qui est de plus en plus utilisée.

Bibliographie

[1] Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J. (2004). *Statistics of Extremes, Theory and Applications*. Wiley. 490 p.

[2] Buck, M.J., Nobel, A.B. & Lieb, J.D. (2005) ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data. *Genome Biology*, 6, R97

[3] Buck, M.J. & Lieb, J.D. (2003) ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, 83, 349–360

[4] 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. & 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, 499–509

[5] Friedman, J.H. (1984) A Variable Span Smoother, Technical Report no. 5, Laboratory for Computational Statistics, Department of Statistics, Stanford University, California.

[6] Glynn, E.F., Megee, P.C., Yu, H.G., Mistrot, C., Unal, E., Koshland, D.E., DeRisi, J.L. & Gerton, J.L. (2004) Genome-Wide Mapping of the Cohesin Complex in the Yeast Saccharomyces cerevisiae. *PLoS Biology*, 2, 9, e259

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

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

[9] Zheng, M., Barrera, L.O., Ren, B. and Wu, Y.N. (2007) ChIP-chip: Data, Model, and Analysis. *Biometrics*, 63, 3, 787–796