Analyse multi-patients de données génomiques

Quentin Grimonprez[1] & Alain Celisse[2] & Guillemette Marot[3]

[1] *DGA & Inria Lille-Nord Europe, quentin.grimonprez@inria.fr*
[2] *Inria Lille-Nord Europe & Laboratoire Paul Painlevé, Université Lille 1, alain.celisse@inria.fr*
[3] *Inria Lille-Nord Europe & EA 2694, Université Lille 2, guillemette.marot@inria.fr*

Résumé. Les technologies à haut-débit en génomique génèrent des observations pour des millions de variables sur des centaines d'individus. Outre le problème de la sélection de marqueurs en très grande dimension qui reste encore très ouvert, des techniques de normalisation et de segmentation des profils génomiques ont été développées ces dernières années. Nous aborderons l'analyse multi-patients de données génomiques de deux manières différentes: (*i*) en fournissant une suite d'outils adaptés à une analyse individuelle complète et en proposant de nouveaux choix de paramètres pour automatiser ces analyses individuelles parfois difficiles (*ii*) en analysant simultanément tous les profils génomiques pour sélectionner des marqueurs qui permettent de prédire la réponse des patients pour une variable donnée (binaire ou quantitative). Nous illustrerons ces deux approches à partir du package `MPAgenomics`, que nous développons sur la forge R.

Mots-clés. statistique et génome, segmentation, SNP, normalisation, nombre de copies, sélection de marqueurs

Abstract. High-throughput technologies generate observations for millions of variables and hundreds of individuals. Aside the burning problem of markers selection in ultra-high dimension, techniques for normalization and segmentation of genomic profiles have been developed over the past few years. We will tackle multi-patients analysis using two different points of views: (*i*) by providing a pipeline for complete individual analysis and suggesting new parameter choices in order to automatize these individual sometimes difficult analyses (*ii*) by analyzing simultaneously all genomic profiles to select markers which would predict patients' response for a given variable (binary or quantitative). We will illustrate these two approaches with the package MPAgenomics, which we are developing on R forge.

Keywords. statistics and genomics, segmentation, SNP, normalization, copy-number, markers selection

1 Introduction

L'analyse de données provenant de puces de génotypage dans R nécessite l'utilisation de plusieurs packages, `aroma` pour la normalisation des puces Affymetrix SNP6.0 [Bengtsson, 2009],

`changepoint` pour la segmentation des profils de nombres de copies [Killick *et al.*, 2013], `cghcall` pour labéliser les segments [van de Wiel *et al.*, 2007], et `glmnet` pour la régression pénalisée [Friedman *et al.*, 2010]. Chaque package effectue une tâche spécifique dans l'analyse mais est indépendant des autres, notamment sur le format des données adopté, rendant l'utilisation conjointe des packages compliquée pour les débutants en R. Une des principales contributions du package R `MPAgenomics` est d'agréger ces packages en fournissant des wrappers pour lier les différentes méthodes automatiquement. Le choix des packages interfacés par `MPAgenomics` a été guidé par l'efficacité des méthodes et la nécessité de gérer des données de grandes tailles. Par exemple, le choix de PELT [Killick *et al.*, 2013] pour l'étape de segmentation dans `MPAgenomics` se base sur la comparaison de différents packages effectuée dans [Hocking *et al.*, 2013]. De plus, `MPAgenomics` propose une nouvelle méthode pour calibrer au mieux le paramètre de la pénalité dans PELT.

Dans la suite, nous décrivons les différentes étapes présentes dans `MPAgenomics` et présentons plus en détails la calibration de paramètre pour la segmentation.

2 Package MPAgenomics

2.1 Normalisation des données

Le processus de normalisation présent dans `MPAgenomics` intègre la *correction de biais techniques* et *l'estimation du nombre de copies et de la fraction d'allèle B* (la fraction d'allèle B correspond à la proportion du signal total provenant de l'allèle B).

L'estimation du nombre total de copies et de la fraction d'allèle B est effectuée par la méthode *CRMAv2* [Bengtsson *et al.*, 2009] implémenté dans la suite `aroma`. La méthode *Tumorboost* [Bengtsson *et al.*, 2010] est aussi proposée pour des études où la contamination entre cellules normales et tumorales est importante.

2.2 Méthode de segmentation et labélisation

Suite aux résultats de [Hocking *et al.*, 2013], nous avons choisi d'utiliser la méthode de segmentation PELT [Killick *et al.*, 2013] dans `MPAgenomics`. Elle se base sur une pénalité de type $\lambda \log(n)$ avec n la longueur du signal et λ un paramètre à choisir. Nous avons observé qu'utiliser la valeur par défaut $\lambda = 1$ sur des données réelles [Renneville *et al.*, 2013] amenait à de la sur-segmentation (trop de segments). Comme le choix de λ est crucial, `MPAgenomics` propose un choix automatique de λ spécifique pour chaque signal (cf Section 3). Une fois les profils du nombre de copies segmentés, la méthode *CGHcall* [van de Wiel *et al.*, 2007] est lancée pour labéliser les différents segments trouvés avec *loss*, *normal* et *gain*.

2.3 Sélection de marqueurs génomiques

Le but de cette étape est de sélectionner des marqueurs génomiques (SNPs ou CNV) associés à une réponse donnée pour l'ensemble des profils.

Pour chaque individu i ($1 \leq i \leq I$), notons y_i la réponse et $x_{i,p}$ la valeur correspondante du nombre total de copies ou de la fraction d'allèle B à la position génomique p ($1 \leq p \leq P$).

En raison du grand nombre de marqueurs ($P \gg I$), `MPAgenomics` utilise la méthode *lasso* [Tibshirani , 1994] afin d'en sélectionner un nombre réduit. La méthode consiste à minimiser $\beta \in \mathbb{R}^P \mapsto g(\beta)$, où

$$g_\rho(\beta) = \sum_{i=1}^{I} (y_i - (X\beta)_i)^2 + \rho \sum_{p=1}^{P} |\beta_p| \ ,$$

avec $(X\beta)_i = \sum_p x_{i,p}\beta_p$ et $\rho > 0$ un paramètre contrôlant le nombre d'éléments non nuls de β.

Les variables sélectionnées sont les plus pertinentes au regard de la réponse.

Dans le cas d'une régression linéaire, `MPAgenomics` fournit efficacement la solution exacte par l'utilisation du nouveau package R `HDPenReg`, lequel est une implémentation de l'algorithme *lars* dédiée à un très grand nombre de marqueurs (plusieurs milliers voire millions). La régression logistique est également disponible dans le cas de réponses binaires par le biais d'une interface du package `glmnet` [Friedman *et al.*, 2010].

La calibration du paramètre de régularisation ρ se fait par validation croisée à K blocs [Arlot et Celisse, 2010].

3 Choix du paramètre de la segmentation

Premièrement, nous détaillons un choix individuel du paramètre λ pour chaque signal que nous proposons avec la méthode de segmentation Pelt. Ensuite, nous illustrons son potentiel comparé à l'utilisation d'un paramètre commun à un ensemble de profils sur données réelles [Renneville *et al.*, 2013].

3.1 Méthode proposée

Pour chaque profil, la méthode Pelt est exécutée sur une grille de valeurs de λ. La figure 1 montre le nombre de segments en fonction de λ sur un exemple.

Le plus long intervalle de λ pour lequel le nombre de segments reste inchangé (et plus grand que 1) indique une grande confiance dans la segmentation obtenue pour ces valeurs. La borne inférieure de cet intervalle est la valeur de λ choisie par `MPAgenomics` pour un signal donné.

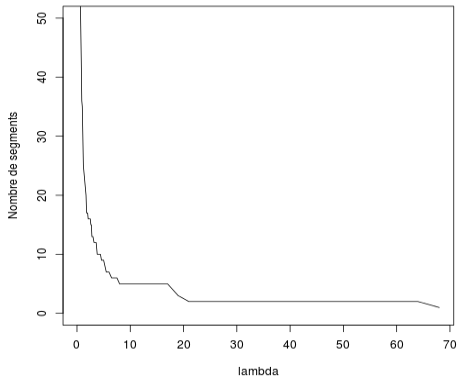

Figure 1: Nombre de segments pour chaque λ dans la pénalité PELT.

3.2 Paramètre commun versus paramètre individuel

Le choix d'un λ spécifique à chaque profil présenté à la section 3.1 est comparé avec l'utilisation d'un λ commun dépendant du ratio signal sur bruit (SNR).

Les profils du jeu de données réelles [Renneville *et al.*, 2013] sont classés en groupes homogènes de SNR par l'utilisation d'un modèle de mélange Gaussien. L'utilisation de trois groupes a été sélectionnée par critère BIC.

La figure 2 montre les résultats (chromosome 1) pour chaque profil (patient) numéroté de 1 à 70. Le plus long intervalle de λ obtenu par la stratégie expliquée en 3.1 est tracé pour chaque patient (dont les numéros sont donnés en ordonnée). Tandis que le

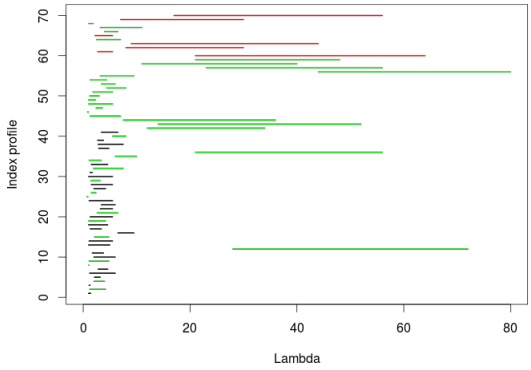

Figure 2: Le plus grand intervalle de λ (axe des abscisses) pour 70 profils du nombre de copies (chromosome 1) (axe des ordonnées). Les couleurs indiquent les classes de ratio signal sur bruit (noir< rouge < vert).

groupe avec le plus faible SNR contient uniquement des faibles valeurs de λ, les autres groupes correspondent à des intervalles de petites et grandes valeurs de λ. Le choix d'un λ commun pour chaque groupe mène à des sur-(sous-)segmentations sur nos données. La

même conclusion s'applique aux autres chromosomes et critères tels que la variance, ce qui justifie l'implémentation du choix de λ spécifique à chaque profil dans `MPAgenomics`.

4 Conclusion

`MPAgenomics` fournit une pipeline facile d'utilisation pour la normalisation et l'analyse multi-patients de données génomiques. Il fournit également des choix automatiques pour le paramètre de segmentation et celui de la sélection de marqueurs. Même si la normalisation est fournie pour les puces Affymetrix, les autres étapes (segmentation, labélisation et sélection de marqueurs) peuvent être appliquées à des données de séquençage à haut débit.

References

[Arlot et Celisse, 2010] Arlot, S. et Celisse, A. (2010), A survey of cross-validation procedures for model selection, *Statistics Surveys*, 4, 40–79.

[Bengtsson, 2009] Bengtsson, H. (2004), aroma - An R Object-oriented Microarray Analysis environment, *Preprint in Mathematical Sciences*.

[Bengtsson *et al.*, 2009] Bengtsson, H. Wirapati, P. et Speed,T.P. (2009), A single-array preprocessing method for estimating full-resolution raw copys from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6, *Bioinformatics*, 25, 2149–2156.

[Bengtsson *et al.*, 2010] Bengtsson, H. *et al* (2010), TumorBoost: Normalization of allele-specific tumor copy numbers from a single pair of tumor-normal genotyping microarrays, *BMC Bioinformatics*, 11, 1–17.

[Efron *et al.*, 2004] Efron, B. *et al* (2004,) Least angle regression, *Annals of Statistics*, 32, 407–499.

[Friedman *et al.*, 2010] Friedman, J. *et al* (2010), Regularization Paths for Generalized Linear Models via Coordinate Descent, *Journal of Statistical Software*, 33, 1-22.

[Hocking *et al.*, 2013] Hocking, T. *et al* (2013), Learning smoothing models of copy number profiles using breakpoint annotations, *BMC Bioinformatics*, 14, 164.

[Killick *et al.*, 2013] Killick, R. et Eckley E. (2013), changepoint: An R package for changepoint analysis, *R package version 1.1*, http://CRAN.R-project.org/package=changepoint.

[Renneville *et al.*, 2013] Renneville, A. *et al* (2013), Clinical impact of gene mutations and lesions detected by SNP-array karyotyping in acute myeloid leukemia patients in the context of gemtuzumab ozogamicin treatment: Results of the ALFA-0701 trial, *Oncotarget*, 4, 9.

[Tibshirani , 1994] Tibshirani, R. (1994), Regression Shrinkage and Selection Via the Lasso, *Journal of the Royal Statistical Society, Series B*, 58, 267-288.

[van de Wiel *et al.*, 2007] van de Wiel, M. *et al* (2007), CGHcall: Calling aberrations for array CGH tumor profiles, *Bioinformatics*, 23, 892-894.