jointSeg : Segmentation de données génomiques en cancérologie

M. Pierre-Jean[a,b] et **P. Neuvial**[b]

[a]EA 2694 Université Lille 2
Centre d'Etudes et de Recherche en Informatique Médicale
1 place de Verdun, 59045 Lille cedex
morgane.pierrejean@genopole.cnrs.fr

[b]UMR 8071 CNRS - Université d'Evry- INRA
Laboratoire Statistique et Génome
23 boulevard de France, 91037 Evry cedex
pierre.neuvial@genopole.cnrs.fr

Mots clefs : Bioinformatique, CNV, Cancer, Nombre de copies, Fraction d'allèle B, biostatistique, détection de ruptures.

L'identification des régions du génome où le nombre de copies d'ADN a été altéré dans les cellules cancéreuses permet de mieux comprendre la progression des tumeurs et de mettre en place des thérapies personnalisées. Nous nous sommes intéressés à la détection de ruptures dans les profils génomiques issus d'échantillons de cellules cancéreuses.

Le package `jointSeg` est disponible depuis janvier 2013 sur R-forge [1]. Il permet notamment :

1. de générer simplement des profils synthétiques réalistes, à l'aide d'un petit nombre de paramètres dont l'interprétation biologique est claire : la proportion de cellules tumorales, la longueur du signal, le nombre de ruptures ;

   ```
   > library(jointSeg)
   > data <- loadCnRegionData(platform="Affymetrix", tumorFraction=.5)
   > set.seed(1)  ## for full reproducibility
   > sim <- getCopyNumberDataByResampling(2e4, nBkp=4, regData=data)
   ```

2. l'utilisation de plusieurs méthodes de segmentation existantes via une interface unifiée :
   – approches exactes par programmation dynamique (`cghseg` [1]) ;
   – segmentation binaire (`CBS` [4], `PSCBS` [6])
   – régression pénalisée de type fused Lasso (`GFLARS` [2], portage en R d'un code Matlab)
   – modèle de Markov caché (`PSCN` [3]).
   Nous avons également implémenté une méthode que nous avons appelée `RBS` pour Recursive Binary Segmentation, et qui combine `CART` et la programmation dynamique [5] :

   ```
   > resRBS <- PSSeg(data=sim$profile, K=20, flavor="RBS", profile=TRUE)
   ```

3. la représentation graphique des résultats (Figure 1) ;

   ```
   > plotSeg(sim$profile, list(true=sim$bkp, est=resRBS$bestBkp))
   ```

4. l'évaluation des performances des différentes méthodes en fonction de la taille d'une zone de tolérance autour des vraies ruptures (non illustrée dans ce résumé pour des raisons de place).

1. `http://r-forge.r-project.org/R/?group_id=1562`

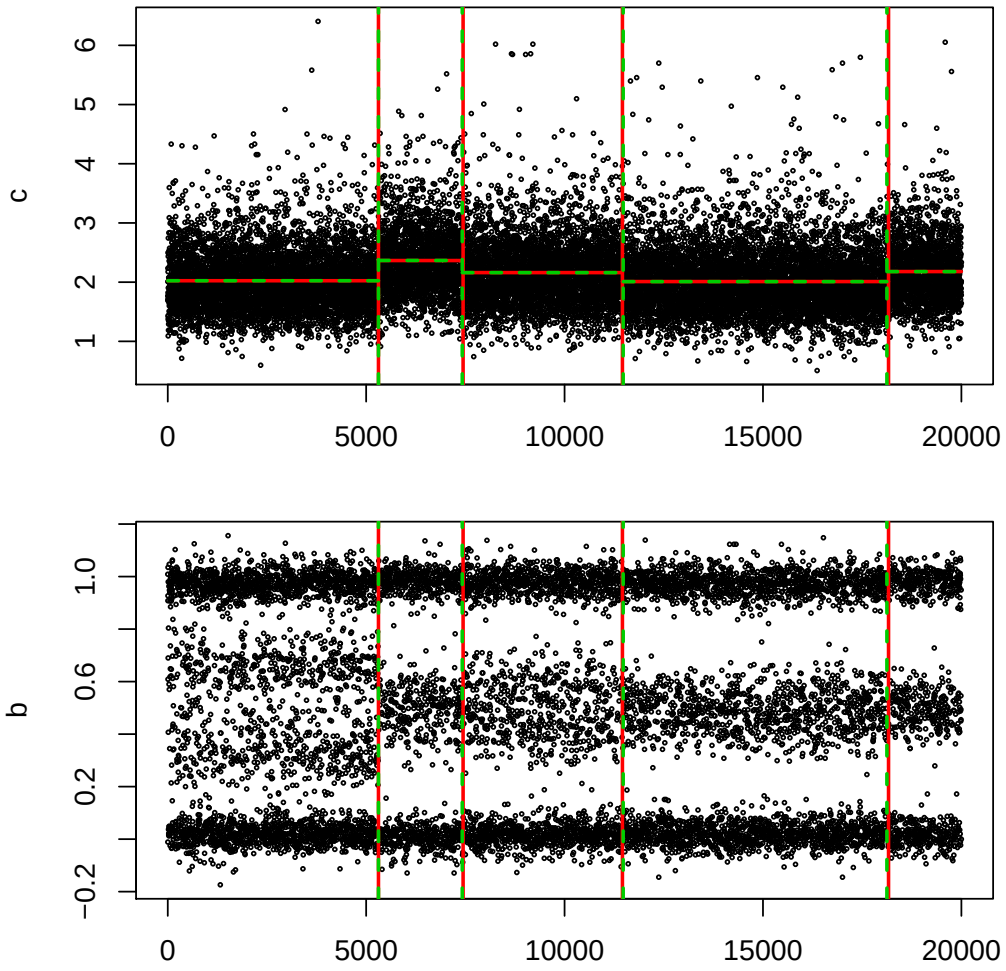

FIGURE 1 – Exemple de données synthétiques produites par le package `jointSeg`. Lignes verticales rouges : vraies ruptures ; lignes verticales vertes : points de ruptures identifiés par `RBS`.

Références

[1] G. Rigaill. Pruned dynamic programming for optimal multiple change-point detection. Technical report, http ://arXiv.org/abs/1004.0887, 2010.

[2] J.-P. Vert and K. Bleakley. Fast detection of multiple change-points shared by many signals using group LARS. *Advances in Neutral Information Processing Systems*, 2010.

[3] Chen, H., Xing, H. and Zhang, N.R. Estimation of parent specific DNA copy number in tumors using high-density genotyping arrays. *PLoS Comput Biol*,2011.

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

[5] Gey, S. and Lebarbier, E., Using CART to Detect Multiple Change Points in the Mean for Large Sample, *Statistics for Systems Biology research group*, (2008)

[6] Olshen, Adam B and Bengtsson, Henrik and Neuvial, Pierre and Spellman, Paul T and Olshen, Richard A and Seshan, Venkatraman E, Parent-specific copy number in paired tumor-normal studies using circular binary segmentation *Boinformatics*, (2011)