Quelques développements statistiques et algorithmiques pour l'analyse de données génomiques

Soutenance d'HDR

G. RIGAILL

IPS2 (Gnet) et LaMME (Stat & Genome)

Septembre 2020

Plan

Structure du mémoire

Bilan comptable et résumé de mes recherches

Chap. 1 Production scientifique

Chap. 2 Tour d'horizon

J'ai essayé de maintenir l'équilibre entre méthodologie et application

4 axes de recherche

Chap. 3 Détection de ruptures multiples

Chap. 4 Analyse de jeux de données omiques

Chap. 5 Classification régularisée

Chap. 6 Évaluation de méthodologies omiques

Structure du mémoire

4 axes de recherche à l'interface

Chapitre 3 - Détection de ruptures multiples

- Présentation d'algorithmes pour maximiser une vraisemblance
- Présentation « plus intuitive » de l'élagage fonctionnel
  - Formalisme des chaînes de Markov cachées
  - Algorithme de Viterbi sur un espace d'états continu

Chapitre 4 - Analyse de jeux de données omiques

- Conduire l'analyse est complexe
  - Des choix de modélisation et de pré-traitement
  - Des choix d'interprétation biologique
  - Interdisciplinarité. . .

Structure du mémoire

4 axes de recherche à l'interface

Chapitre 5 - Classification régularisée

- Je décris quelques contributions méthodologiques
- Idées : relaxation convexe et intégration de classifications

Chapitre 6 - Évaluation de méthodologies omiques

- Il n'est pas facile de choisir/justifier une stratégie d'analyse
- Je décris quelques travaux visant à éclairer ce choix
- Idées : annotations, ré-échantillonage

Quelques défis pour analyser des données omiques

Dans cette présentation

Développer des méthodes

- Permettre l'inférence de modèles de plus en plus complexes
- Améliorer l'inférence de modèles existants

Chap 3 Évolution des algorithmes d'élagage fonctionnel

Travailler à la mise en œuvre des méthodes

- L'interprétation biologique du modèle n'est pas évidente

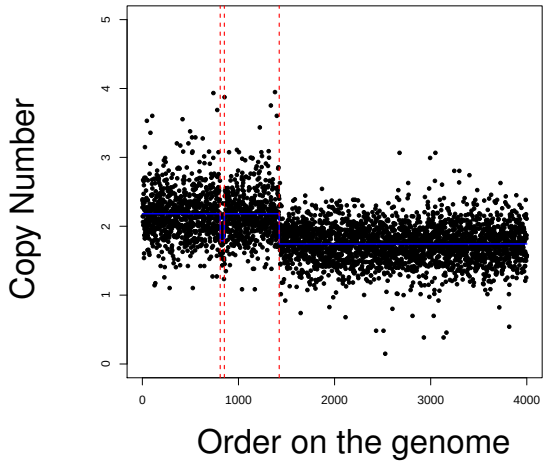
Chap 4 Ruptures pour le transcriptome

Plan

Un problème important et récurrent

Selon le « National Research Council (US) »

- Des modèles complexes
  - Genomique [Hocking *et al.* 2016, Pierre-Jean *et al.* 2015...]
  - Geologie, Finance, Biologie...



- De plus en plus de méthodes ces dix dernières années
  - Modèle gaussien univarié : [Harchaoui and Levy-Leduc 2009, Killick *et al.* 2011, Frick *et al.* 2014, Lin et al. 2015, Dette and Wied 2015, Haynes *et al.* 2016, Maidstone *et al.* 2017, Fryzlewicz 2017...]

Méthodes minimisant un coût pénalisé

- De bonnes propriétés statistiques [Yao 1989, Lebarbier 2005, Arlot *et al.* 2012...]
- Complexité algorithmique quadratique $\mathcal{O}(n^2)$ ou moins
- Etat de l'art pour plusieurs applications en biologie

Un modèle constant par morceaux

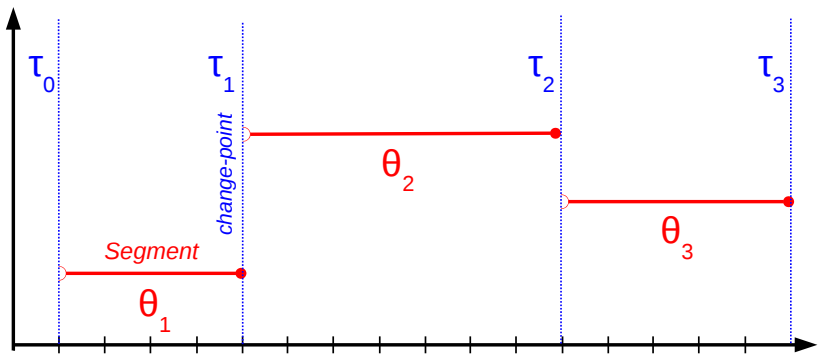
données $Y_1, \ldots, Y_n$

ruptures $\tau = (\tau_1, .., \tau_D)$

segments $s_d = (\tau_{d-1}, \tau_d]$

paramètres $\theta = (\theta_1, .., \theta_D)$

modèle $Y_i \sim \mathcal{F}(\theta_d)$ i.i.d



Modèle $\mathcal{F}$ $\rightarrow$ γ $\rightarrow$ Optimisation

$Y_i \sim \mathcal{N}(\theta_d, \sigma^2)$ $\quad (Y_i - \theta_d)^2$ $\quad \sum_{d=1}^{|\tau|} \sum_{\tau_{d-1}+1}^{\tau_d} (Y_i - \hat{\theta}_d)^2$

- le nombre et la position des ruptures sont inconnus
- l'ensemble des segmentations $\mathcal{M}_n$ est grand : 2^{n-1}

Maximum de vraisemblance pénalisée

- Minimiser moins la log-vraisemblance

$$cost(\tau) = \sum_{d=1}^{|\tau|} \min_{\theta} \{ \sum_{\tau_{d-1}+1}^{\tau_d} \gamma(Y_i, \theta) \}$$

- $cost(\tau)$ est minimal pour $n-1$ ruptures.
- Il faut pénaliser !

$$cost(\tau) = \sum_{d=1}^{|\tau|} \min_{\theta} \{ \sum_{\tau_{d-1}+1}^{\tau_d} \gamma(Y_i, \theta) \quad + \lambda \}$$

Pénalités linéaires et concaves

- (P1) Linéaire [Yao 1989] ou (P2) Concave [Lebarbier 2005]

$$
\begin{aligned}
P1 : \quad & pen(\tau) &&= 2\sigma^2 |\tau| \log(n) \\
P2 : \quad & pen(\tau) &&= 2\sigma^2 |\tau| (c_1 \log(n/|\tau|) + c_2)
\end{aligned}
$$

Algorithmiquement

- Résoudre (P2) se ramène à résoudre plusieurs problèmes (P1)

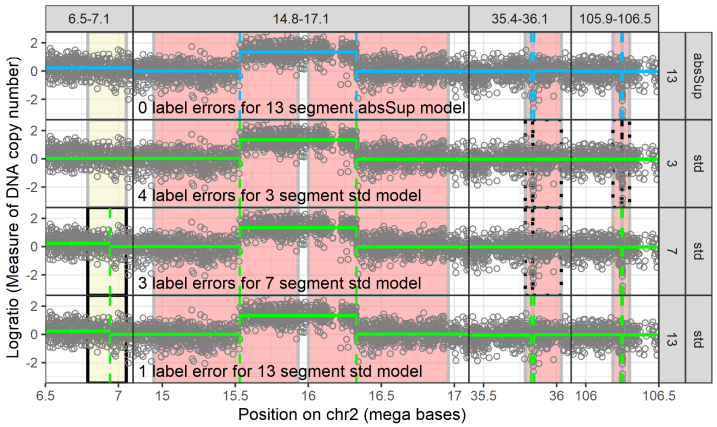
  [Killick *et al.* 2012, Haynes *et al.* 2016]

Plan

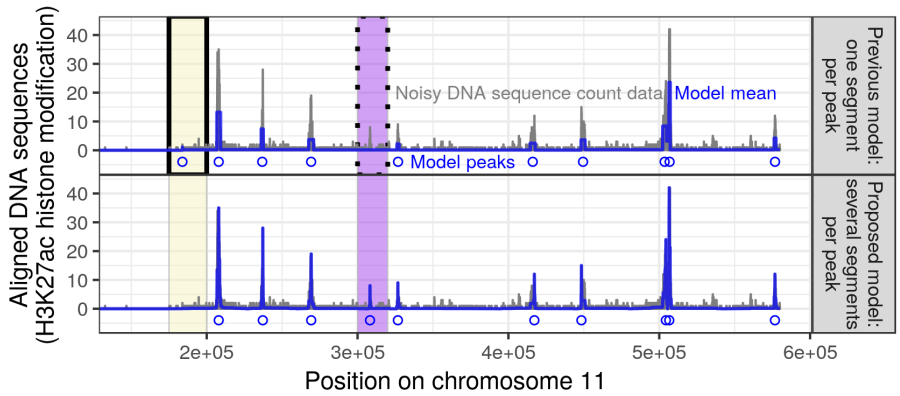
Récurrence fonctionnelle

Prendre en compte des contraintes et des dépendances

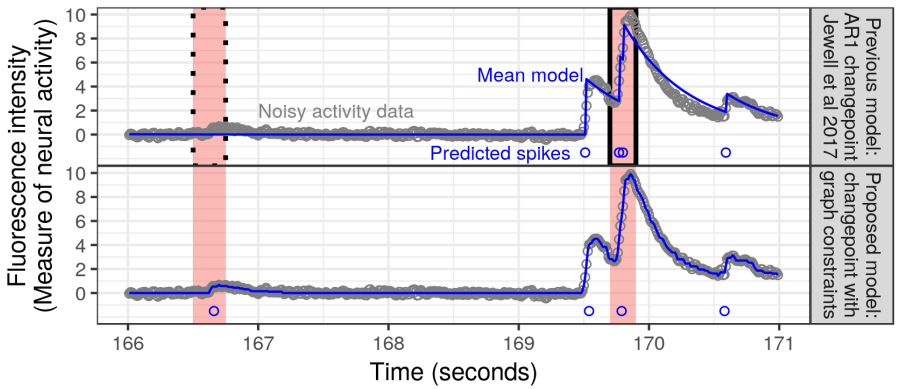
Nombre de copies d'ADN



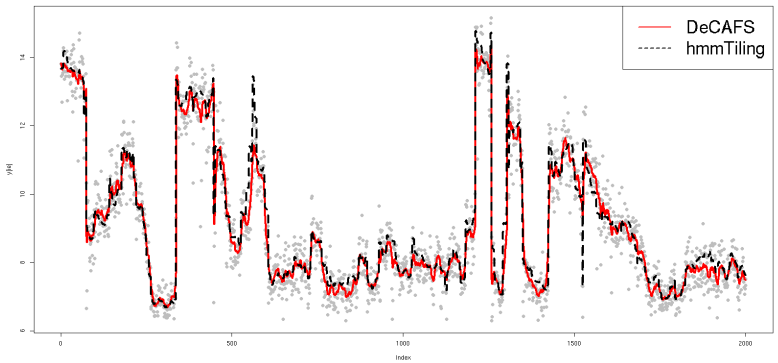
Chip-Seq



Trains de potentiels d'action

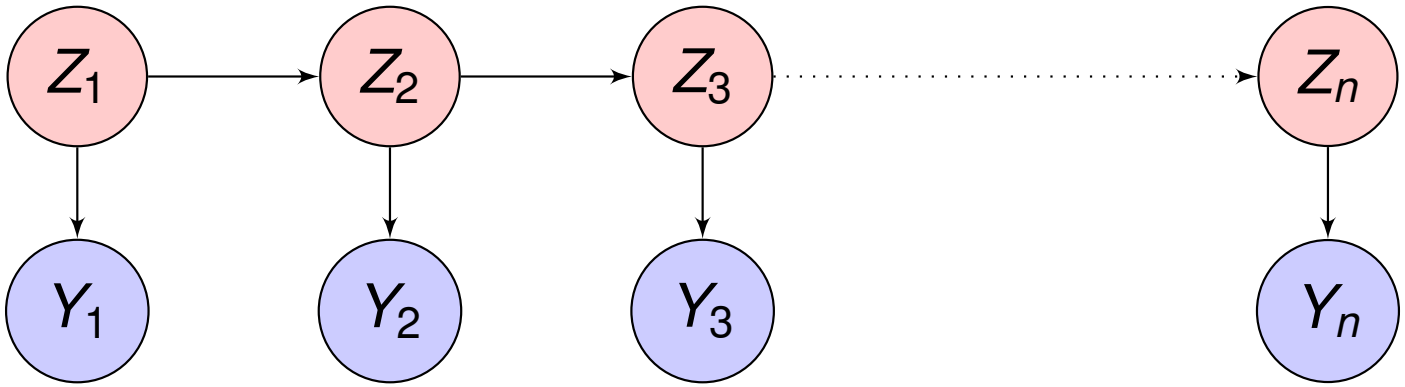


Transcriptome bactérien



Plan

Ruptures et chaînes de Markov cachées



Espace d'états continu Z_i dans un intervalle de $\mathbb{R}$

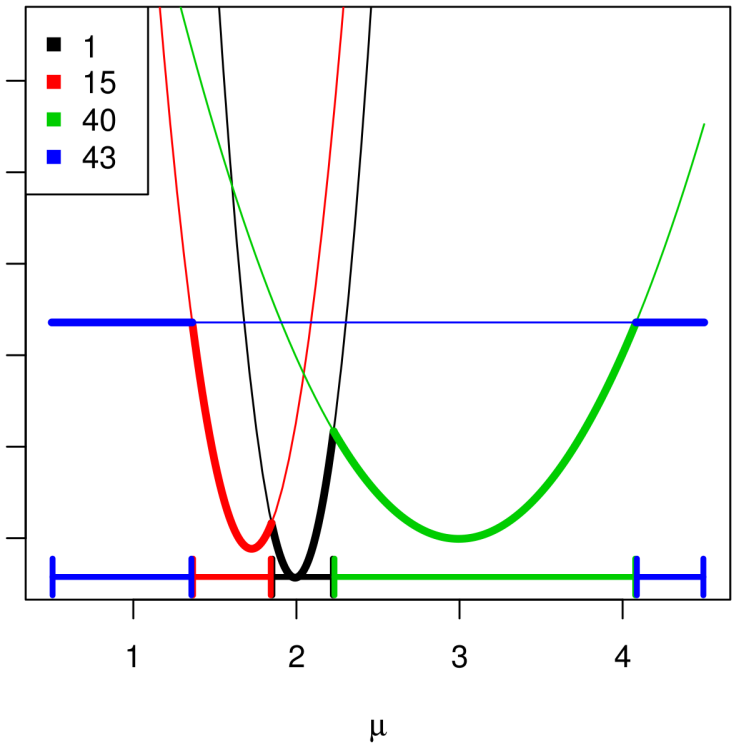
Règle de chaînage $(Y_i|Z_i = \theta) \sim \mathcal{N}(\theta, \sigma^2)$

Transition $k(x, y) \propto \mathbb{I}_{x=y} + e^{-\lambda}\mathbb{I}_{x\neq y}.$

Conditionner par le paramètre du dernier segment

Un point de vue fonctionnel [Rigaill 2010-2015, Johsnon 2010-2013, Rote 2012, Maidstone *et al.* 2016]

$$\widetilde{P}_n(\mu) = \min_{\substack{\tau,\theta \\ \theta_{|\tau|}=\mu}} \left\{ \lambda|\tau| + \sum_{d=1}^{|\tau|} \sum_{\tau_{d-1}+1}^{\tau_d} (Y_i - \theta_d)^2 \right\}$$



Algorithme de Viterbi sur un espace d'états continu

Élagage fonctionnel

$$\widetilde{P_{n+1}}(\mu) = \min \left\{ \begin{array}{c} \text{“pas de rupture”} \\ \widetilde{P}_n(\mu) \\ \\ \min_{\mu'} \{\widetilde{P}_n(\mu')\} + \lambda \\ \text{“une rupture”} \end{array} \right\} + (Y_{n+1} - \mu)^2$$

- L'espace d'états est continu...
- On applique la mise à jour par intervalle
- Au pire il y a $2n - 1$ intervalles.
- Complexité au pire en $\mathcal{O}(n^2)$

Mise à jour fonctionnelle visuellement

$n = 43$

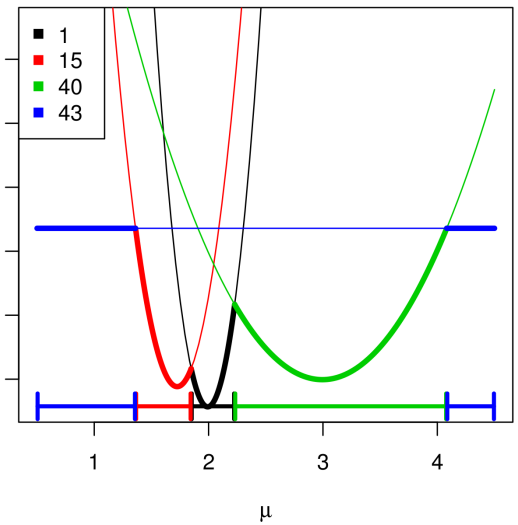
$n = 43 + \frac{1}{2}$

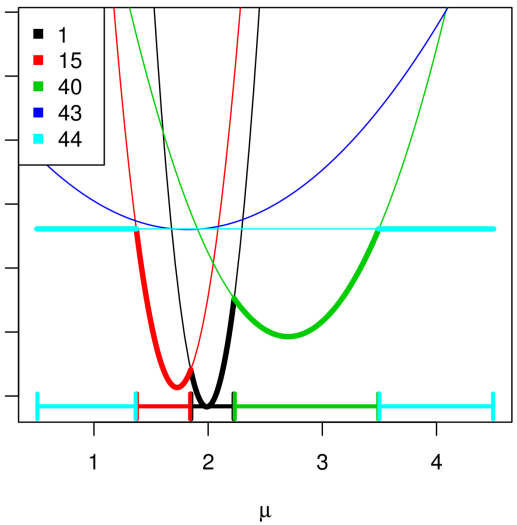
$+(Y_{44} - \mu)^2$

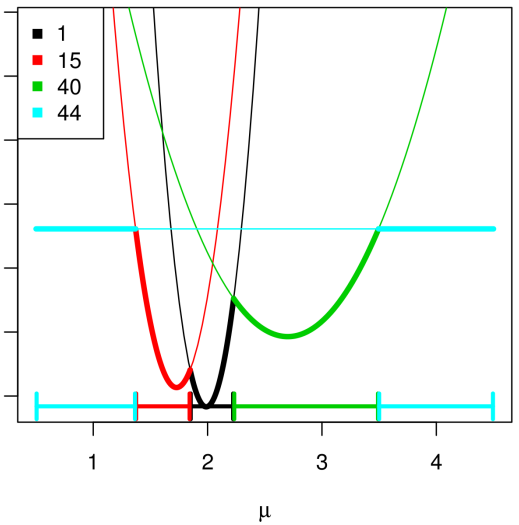
vs. $P_{44} + \lambda$

$n = 44$

discard $\tau = 43$



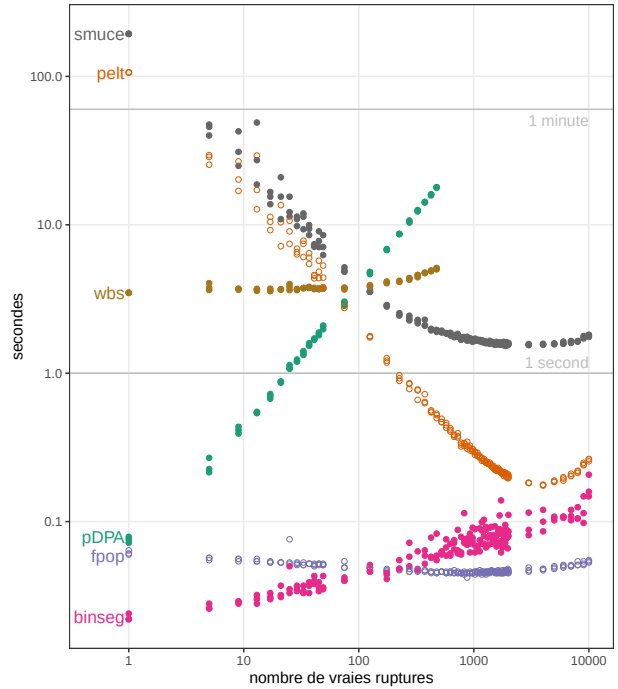




Algorithme Fpop en pratique

- Quasi-linéaire en moyenne *(même sans rupture)*
- Environ 4 secondes pour $n = 10^7$

Pour $n = 10^5$ et D^*



Rapide avec d'excellentes performances statistiques

- Bonnes propriétés statistiques [Yao 1989, ..., Garreau and Arlot 2017]
- SOTA sur des simulations [Fryzlewicz 2014, Fearnhead and Rigaill 2020]
- SOTA pour le nombre de copies d'ADN [Hocking *et al.* 2013, Pierre-Jean *et al.* 2015]

Plan

Extensions de l'élagage fonctionnel

Modèles i.i.d

- Poisson, Binomial Negatif [Cleynen *et al.* 2015]
- Pertes robustes (pour les outliers) [Fearnhead and Rigaill 2018]
- Pas efficace pour des modèles multivariés ?

Modèles avec dépendances

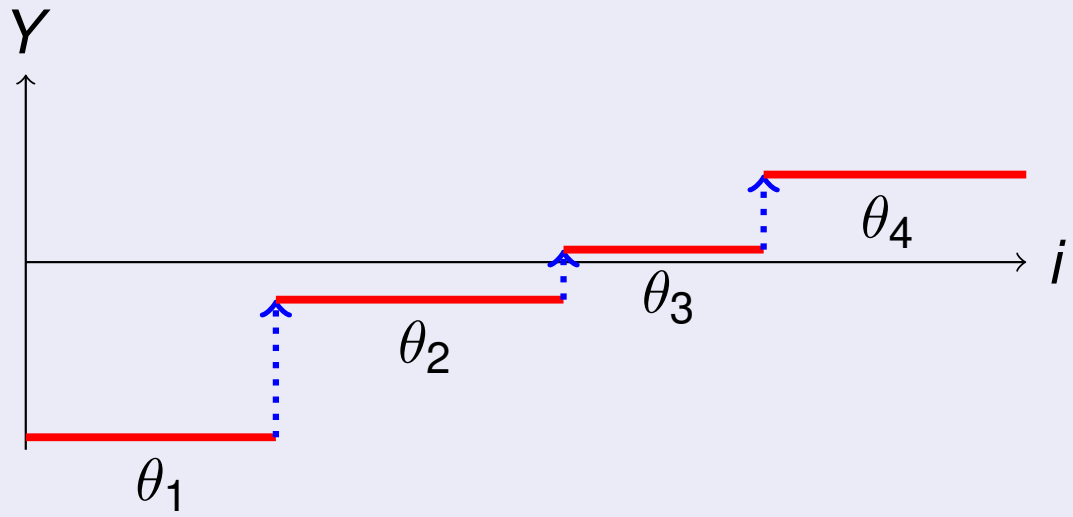
[Maidstone *et al.* 2017, Hocking *et al.* 2018, Jewell *et al.* 2019, Romano *et al.* 2020]

- Isotonique, pics, train de potentiels d'action
- Auto-corrélation, drift ...

Modèle isotonique

Quelques opérateurs [Hocking *et al.* 2018]

$$\theta_1 \leqslant \theta_2 \leqslant \theta_3 \dots$$



$$\widetilde{P}^{Iso}_{n+1}(\mu) = \min \left\{ \begin{array}{c} \text{“pas de rupture”} \\ \widetilde{P}^{Iso}_{n}(\mu) \\ \\ \min\limits_{\mu' \leqslant \mu} \{\widetilde{P}^{Iso}_{n}(\mu')\} + \lambda \\ \text{“une rupture”} \end{array} \right\} + (Y_{n+1} - \mu)^2$$

Opérateur minimum à gauche

Définition

$$\widetilde{min}(f)(\mu) = \min_{\mu' \leqslant \mu} f(\mu').$$

Propriétés

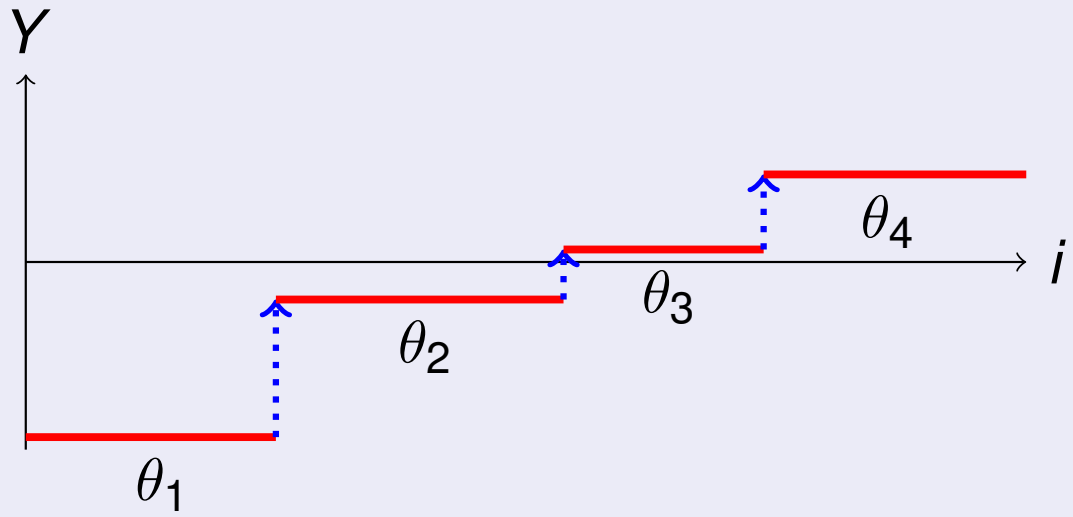
Si f est quadratique sur m intervalles :

1. $\widetilde{min}(f)$ est quadratique sur au plus $2m$ intervalles
2. On peut calculer $\widetilde{min}(f)$ de gauche à droite en $O(m)$

Modèle isotonique

Quelques opérateurs [Hocking *et al.* 2018]

$\theta_1 \leqslant \theta_2 \leqslant \theta_3 \ldots$



$$\widetilde{P}^{Iso}_{n+1}(\mu) = \min \left\{ \begin{array}{c} \text{“pas de rupture”} \\ \widetilde{P}^{Iso}_{n}(\mu) \\ \\ \widetilde{mn}(\widetilde{P}^{Iso}_{n})(\mu) + \lambda \\ \text{“une rupture”} \end{array} \right\} + (Y_{n+1} - \mu)^2$$

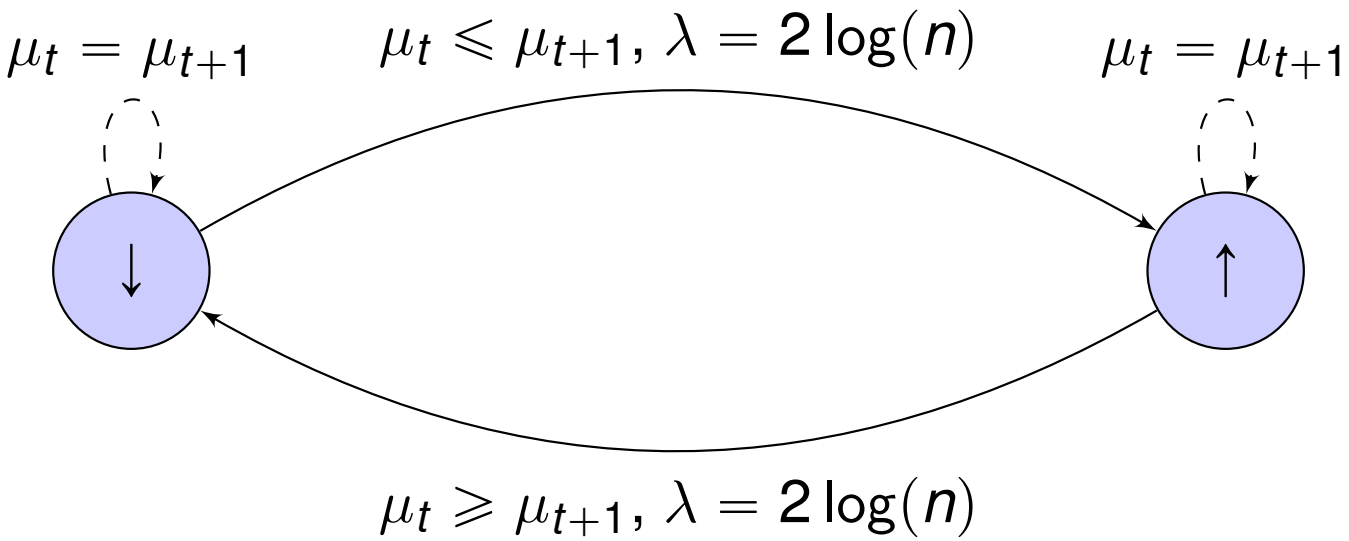
Pour des patterns plus complexes

On considère plusieurs fonctions à chaque étape :

$$s \in \mathcal{S}, \quad \widetilde{P}_n^s$$

- $\mathcal{S}$ est un ensemble fini
- L'espace d'états est $\mathcal{S} \times \mathbb{R}$
- Un noyau de transition représenté par un graphe $\mathcal{G}$

Noyau et graphe pour modéliser des pics



Récurrence fonctionnelle

On obtient $\widetilde{P}^{s}_{n+1}$ en appliquant des opérateurs et en comparant les fonctions de l'étape n.

- Pour $\widetilde{P}^{\uparrow}_{n+1}$ on utilise l'opérateur $\widetilde{mn}$.

$$\widetilde{P}^{\uparrow}_{n+1}(\mu) = \min \left\{ \begin{array}{c} \text{“pas de rupture”} \\ \widetilde{P}^{\uparrow}_{n}(\mu) \\ \\ \widetilde{mn}(\widetilde{P}^{\downarrow}_{n}(\mu)) + \lambda \\ \text{“une rupture”} \end{array} \right\} + (Y_n - \mu)^2.$$

- Pour $\widetilde{P}^{\downarrow}_{n+1}$ on applique l'opérateur minimum à droite...

Un package pour un graphe générique

Thanks to Vincent Runge [Runge et al. submitted]

Algorithme gfpop

- En entrée : un graphe, un(e) modèle/distribution et des données
- Applications
  - Chip-Seq et pics [Hocking et al. 2017 and 2018]
  - Potentiels d'actions [Jewell et al. 2019]

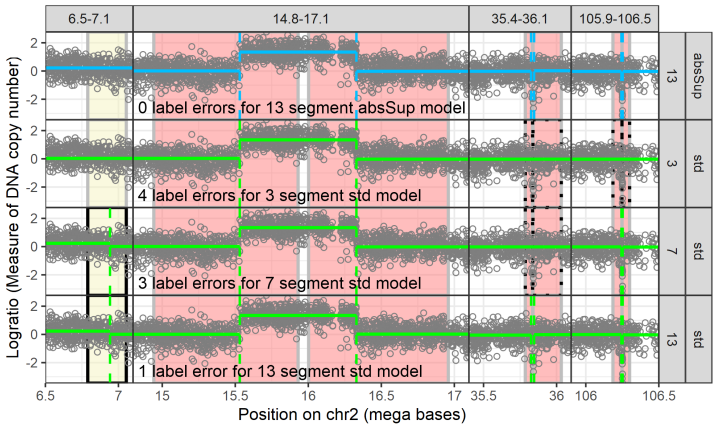
Pattern	perte	n	temps
Pics	ℓ_2	10^6	$\sim 13s$
Pics	robuste	10^6	$\sim 40s$

https ://github.com/vrunge/gfpop

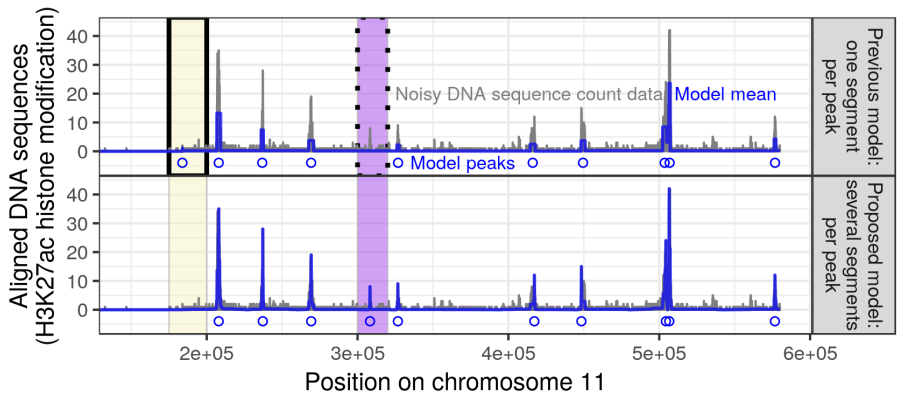
Récurrence fonctionnelle

Prendre en compte des contraintes

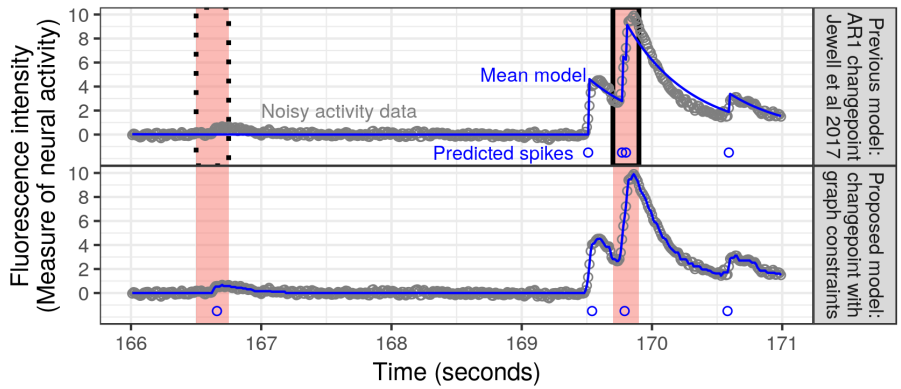
Amplitude des sauts



Des pics



Isotonique Haut-Bas

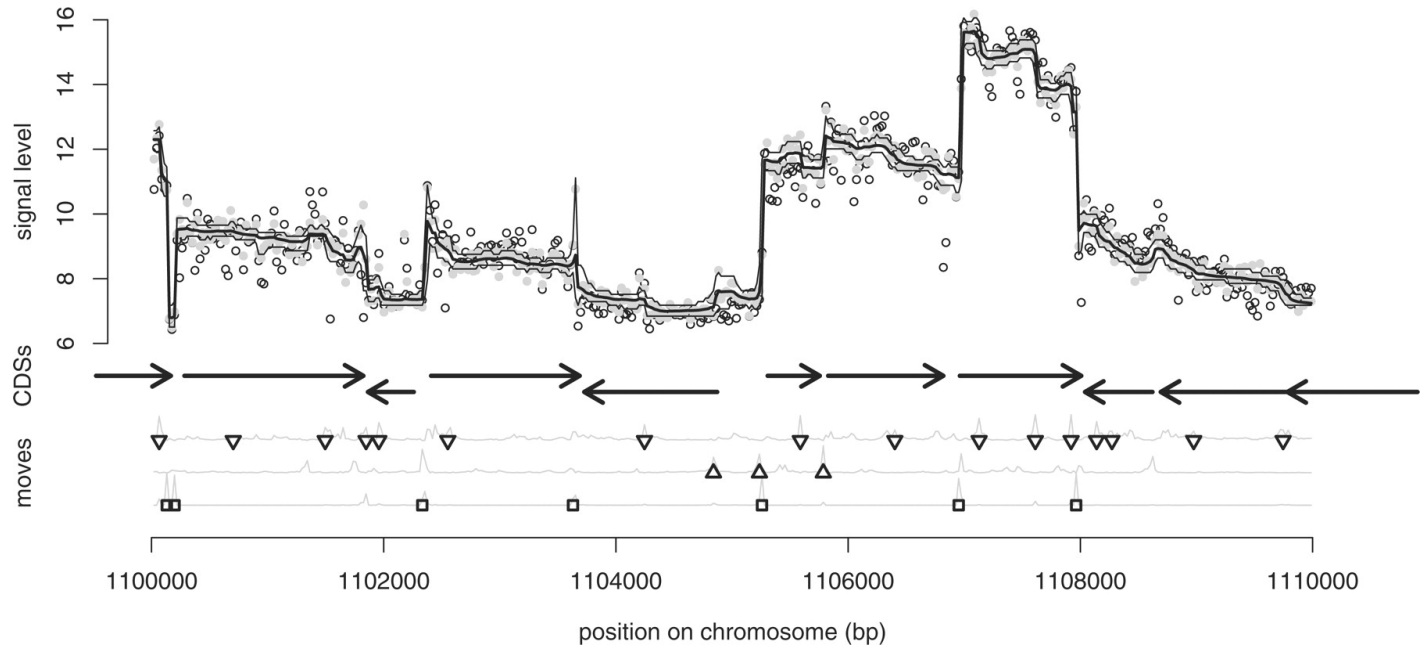


Plan

Plan

Niveau d'expression de cellules bactériennes

HMMTiling [Nicolas et al. 2009]



- Puces tiling avec une résolution inférieure à 25 paires de bases
- Pour chaque sonde t une mesure d'expression : Y_t
- Un mélange de sauts et de variations plus continues...

Modèle de détection de ruptures

avec un drift et de l'auto-corrélation

$$
\begin{aligned}
Y_t &= \theta_t + \epsilon_t \\
\theta_t &= \theta_{t-1} + \eta_t + \delta_t \qquad \text{marche ou saute} \\
\epsilon_t &= \phi\epsilon_{t-1} + \nu_t \qquad \text{AR(1)}
\end{aligned}
$$

Marche aléatoire ou saut

- Marche aléatoire : $\eta_t \sim \mathcal{N}(0, \sigma_\eta^2),\ \delta_t \in \mathbb{R}$
- Saut de δ_t et pénalité λ : si $\delta_t \neq 0$ alors $\eta_t = 0$

Auto-corrélation

- AR(1) : $\nu_t \sim \mathcal{N}(0, \sigma_\nu^2)$

Quelques opérateurs de plus

DeCAFS [Romano et al. 2020]

- Le coût $\widetilde{P}_n^{\#}$ est quadratique par morceaux
- Deux opérateurs de convolution calculables en $O(m)$

DAC : drift + auto-corrélation
SAC : saut + auto-corrélation

$$\widetilde{P}_{n+1}^{\#}(\mu) = \min \left\{ \begin{array}{c} \text{“drift et auto-corrélation”} \\ DAC(\widetilde{P}_n^{\#}(\mu)) \\ \\ SAC(\widetilde{P}_n^{\#}(\mu)) + \lambda \\ \text{“saut et auto-corrélation”} \end{array} \right\} + (Y_n - \mu)^2.$$

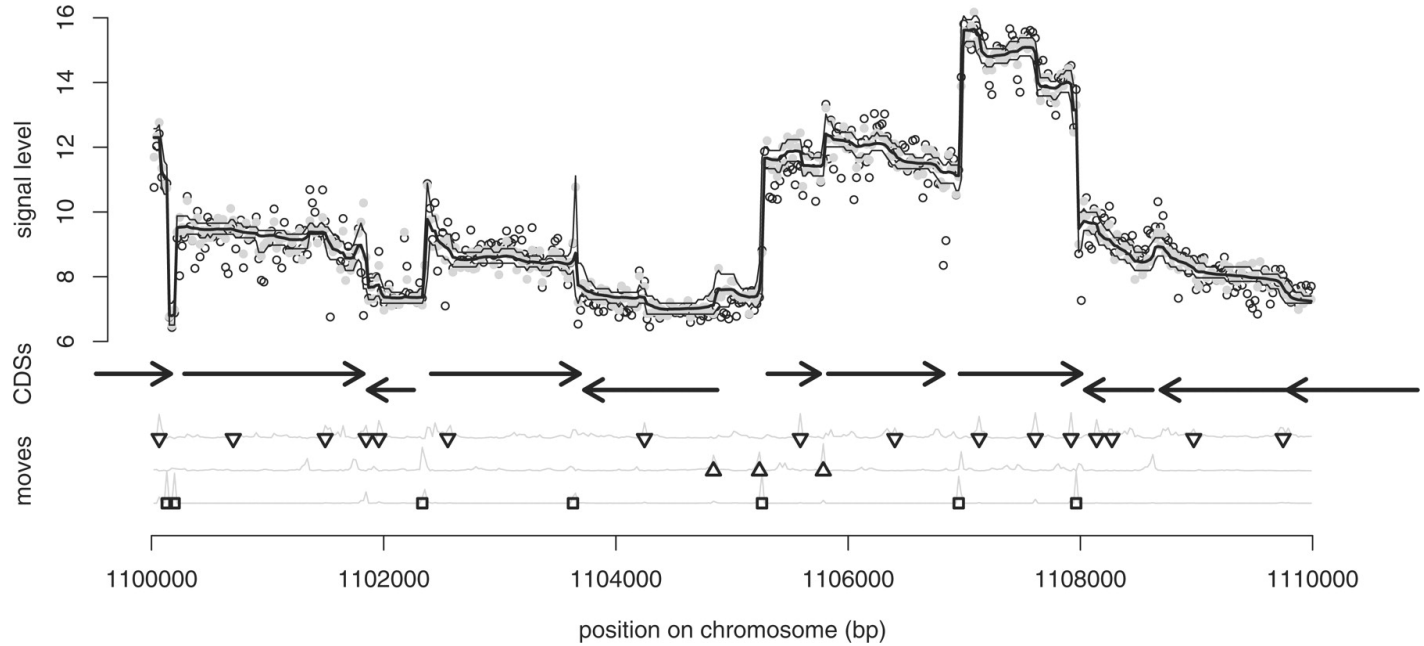
Temps de calcul

Quelques minutes pour $n = 300\,000$

Plan

Prédire des terminateurs et des promoteurs

Première interprétation



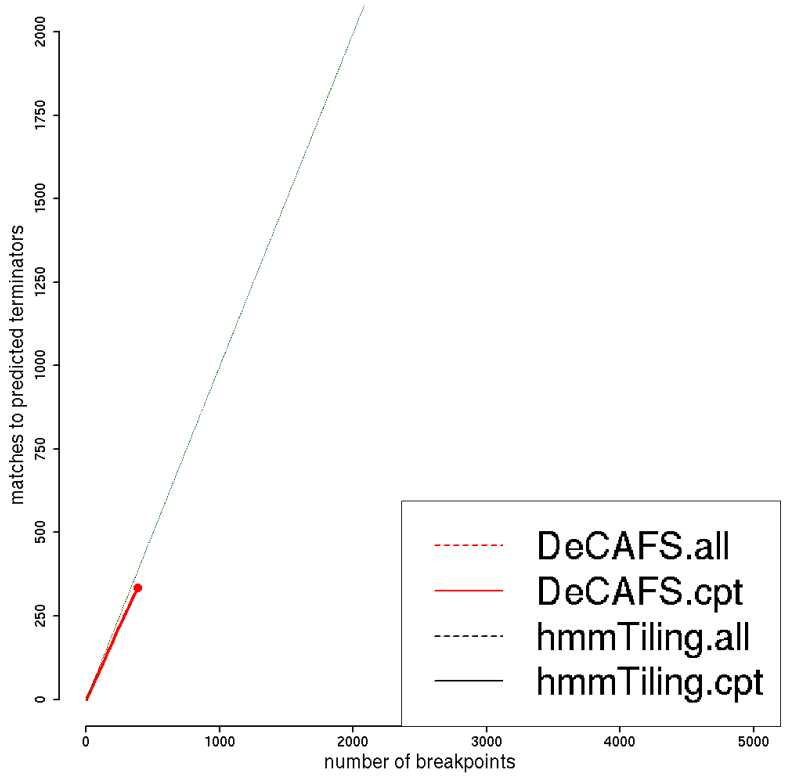
- Un terminateur correspond à la fin d'un gène
- On attend une baisse du niveau de transcription [Nicolas et al. 2009]
  - Un saut : $t \in \hat{\tau}$
  - Et une différence importante : $d_t = \hat{Y}_{t+1} - \hat{Y}_t$ très négative

Interpréter les ruptures ?

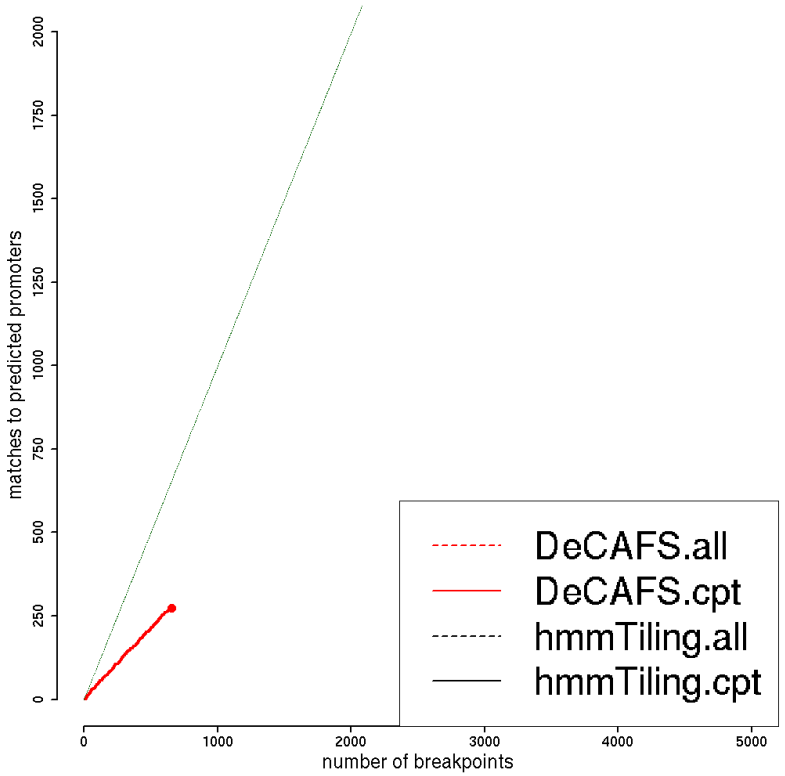
Comparer aux terminateurs connus bioinformatiquement

- axe des y : nombre de terminateurs connus
- axe des x : nombre de terminateurs prédits

Terminateurs

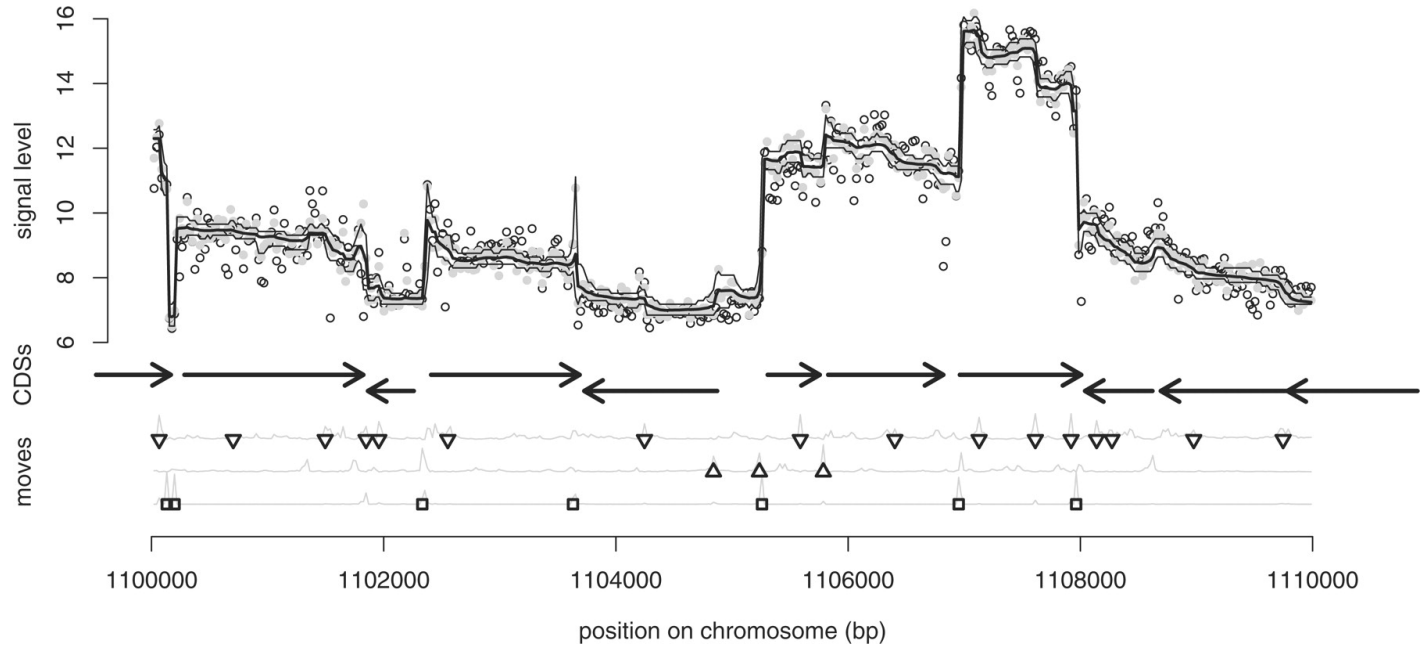


Promoteurs



Prédire des terminateurs et des promoteurs

Deuxième interprétation



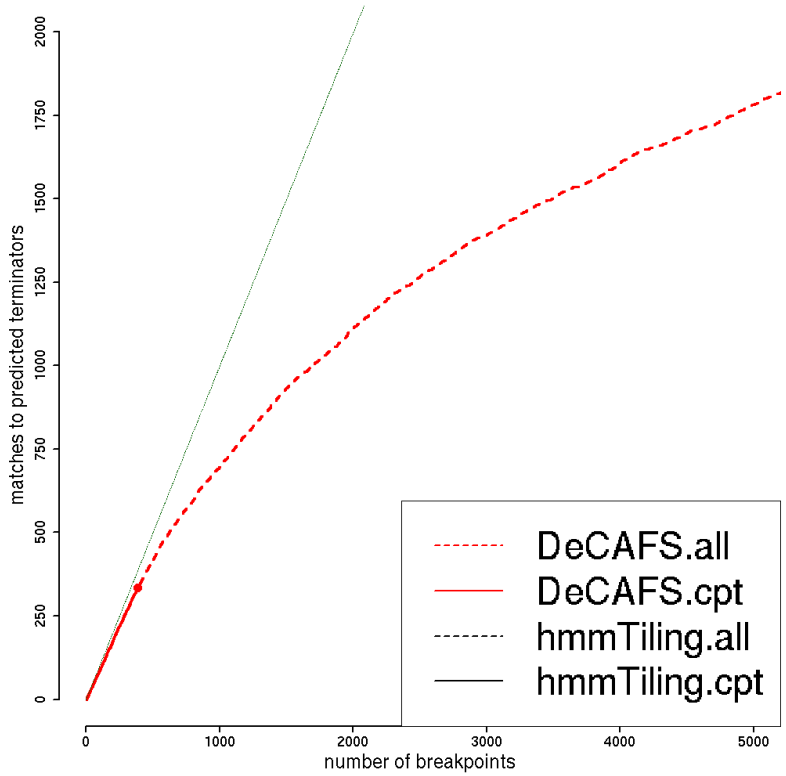
- Un terminateur correspond à la fin d'un gène
- On attend une baisse du niveau de transcription
  - $t \in \{1, \cdots, n\}$
  - Une différence importante : $d_t = \hat{Y}_{t+1} - \hat{Y}_t$ très négative

Interpréter les différences ?

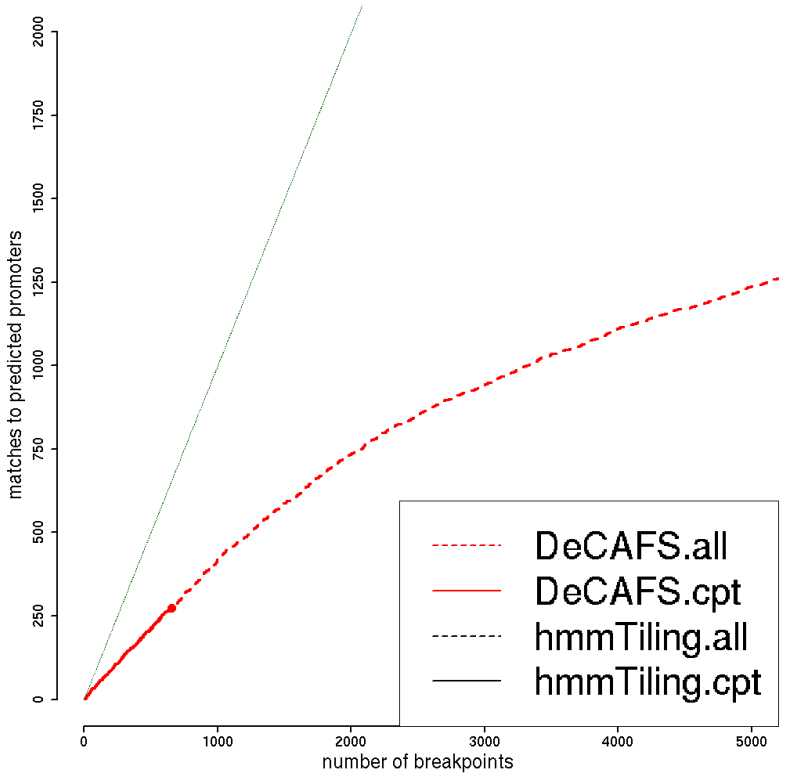
Comparer aux terminateurs connus bioinformatiquement

- axe des y : nombre de terminateurs connus
- axe des x : nombre de terminateurs prédits

Terminateurs



Promoteurs

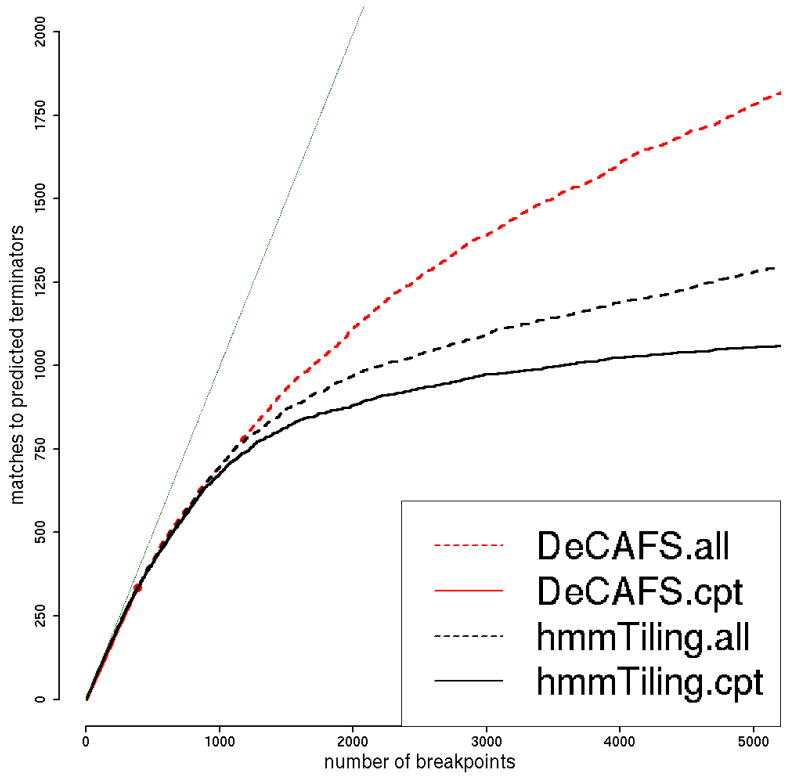


Même conclusion avec HMMTiling

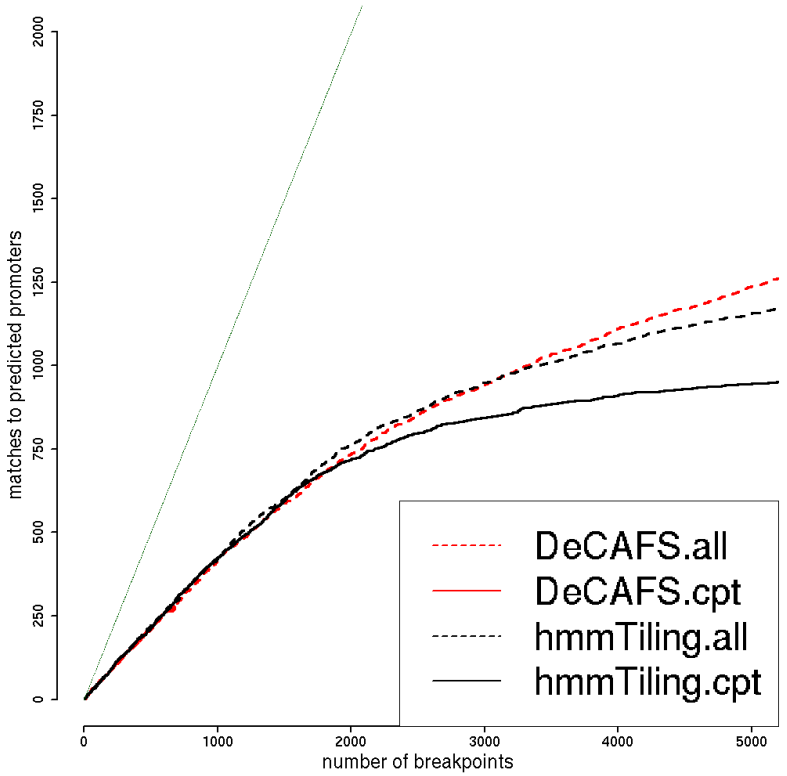
Special Thanks to Pierre Nicolas !

- Il semble préférable de considérer les différences
- Les sauts seuls, $\hat{\tau}$, ne capturent pas idéalement la « réalité »

Terminateurs



Promoteurs



Conclusion

Avec le même modèle deux résultats différents

- Modéliser de petites transitions et des sauts [Nicolas et al. 2009]
- Les sauts seuls, $\hat{\tau}$, ne capturent pas idéalement la « réalité »
- Il semble préférable de regarder : $\hat{Y}_{t+1} - \hat{Y}_t$
- Comment choisir sans informations extérieures ?
- Entre modélisation statistique et interprétation biologique il y a encore beaucoup de travail

Plan

Quelques défis pour analyser des données omiques

Développer des méthodes *Illustré avec le Chap. 3 & 5*

- Permettre l'inférence de modèles de plus en plus complexes
- Améliorer l'inférence de modèles existants

Travailler à la mise en œuvre des méthodes *Illustré avec le Chap. 4*

- Réfléchir au pré-traitement, à la modélisation et à l'interprétation
- Nécessite un dialogue interdisciplinaire

Evaluer *Illustré avec le Chap. 6*

- Toute analyse implique de nombreux choix
  - Pré-traitement, modélisation, méthode, interprétation...
- Il est souvent difficile de juger et de justifier ces choix

Évaluer ce n'est pas facile !

Statistiquement

- Étude mathématique ou de simulations
- On peut explorer de nombreuses configurations
- Ces simulations ne peuvent rivaliser avec la complexité des données biologiques
- Il est douteux qu'une méthode qui fonctionne mal statistiquement soit performante sur des données biologiques

Biologiquement

- Valider expérimentalement les résultats
- Il n'y a pas à douter du réalisme des données
- Les validations sont souvent peu nombreuses et nous pouvons douter de leur représentativité

Quelques défis pour analyser des données omiques

Développer des méthodes

Illustré avec le Chap. 3 & 5

- Permettre l'inférence de modèles de plus en plus complexes
- Améliorer l'inférence de modèles existants

Travailler à la mise en œuvre des méthodes

Illustré avec le Chap. 4

- Réfléchir au pré-traitement, à la modélisation et à l'interprétation
- Nécessite un dialogue interdisciplinaire

Evaluer

Illustré avec le Chap. 6

- Toute analyse implique de nombreux choix
  - Pré-traitement, modélisation, méthode, interprétation...
- Il est souvent difficile de juger et de justifier ces choix

Merci à tous. . .

- Membres du jury
- Collègues et collaborateurs
- Amis et famille