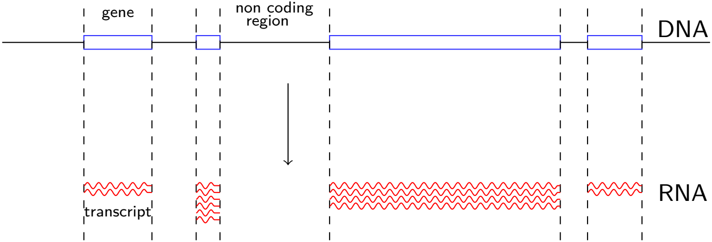
Détection de ruptures dans les données de séquençage

Alice Cleynen

CNRS - Université de Montpellier

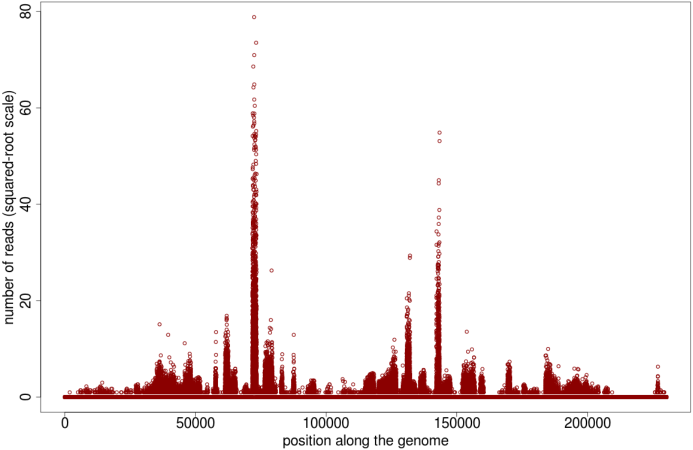
5 Janvier 2017

From DNA to RNA

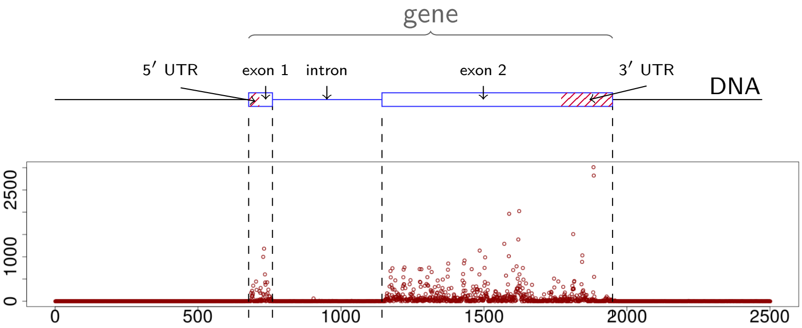


RNA-Seq data

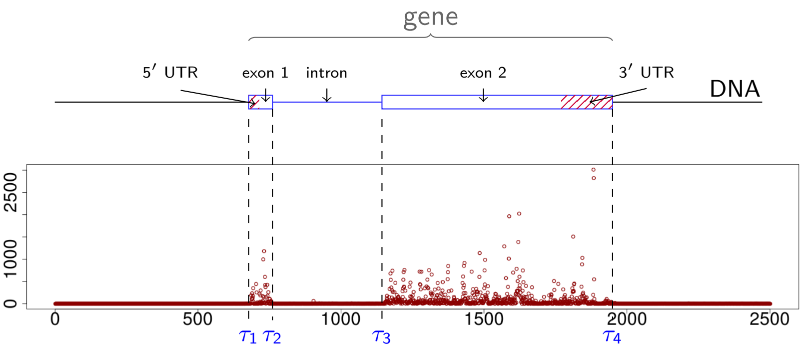
Mapping to the genome



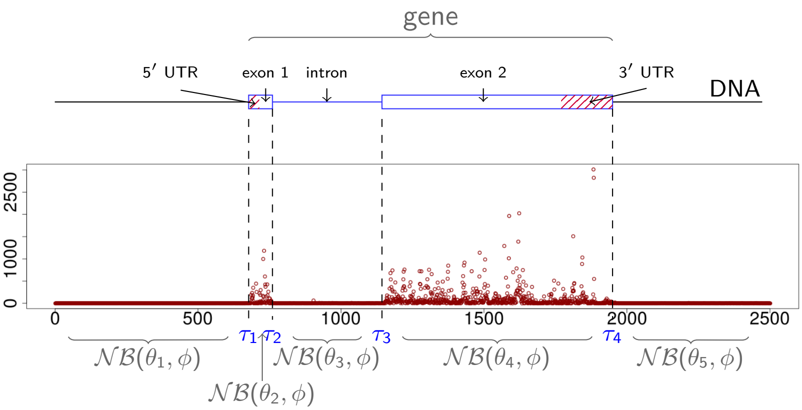
Gene structure



Segmentation Framework

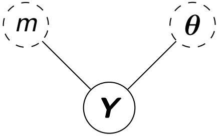


Segmentation Framework



Segmentation Framework

- m a partition of $\{1, \ldots, n\}$
  - $n =$ length of profile
  - $K =$ number of segments
  - $m = \{\tau_1, \ldots, \tau_{K-1}\}$
  - J a segment of m
- $\boldsymbol{\theta} =$ probability parameter of $\mathcal{NB}$
  - $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_K\}$
  - $\phi =$ overdispersion (known)
- $\forall J \in m, \forall t \in J, Y_t \sim \mathcal{NB}(\theta_J, \phi)$



Optimization framework

Optimal segmentation in K segments ?

- Cost of a segmentation m: $\ell(\mathbf{Y}, m)$
- To be minimized over $\mathcal{M}_K$ of cardinality $|\mathcal{M}_K| = \binom{n-1}{K-1}$
  $\rightarrow$ exhaustive search intractable

Optimization framework

Optimal segmentation in K segments ?

- Cost of a segmentation m: $\ell(\mathbf{Y}, m)$
- To be minimized over $\mathcal{M}_K$ of cardinality $|\mathcal{M}_K| = \binom{n-1}{K-1}$
  $\rightarrow$ exhaustive search intractable

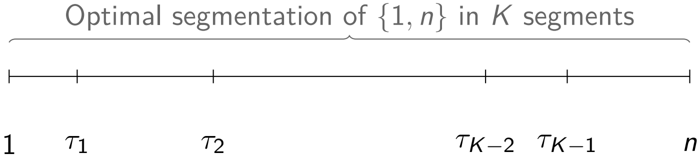
Under a summation assumption:

$$\ell(\mathbf{Y}|m, \theta) = \sum_{r \in m} \ell(Y^r|\theta_r)$$

$\rightarrow$ Dynamic programming algorithm (DPA) provides optimal segmentation in $\mathcal{O}(Kn^2)$

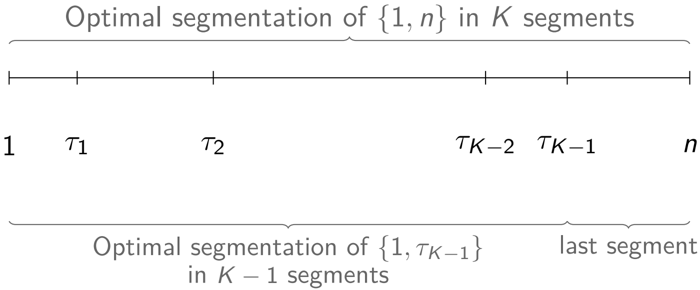
DPA : Bellman (1961); Auger and Lawrence (1989)

How it works



Optimal segmentation of $\{1, n\}$ in K segments
1
τ_1
τ_2
τ_{K-2}
τ_{K-1}
n

How it works



Optimal segmentation of $\{1, n\}$ in K segments
1
τ_1
τ_2
τ_{K-2}
τ_{K-1}
n
Optimal segmentation of $\{1, \tau_{K-1}\}$
in $K-1$ segments
last segment

How it works (with equations)

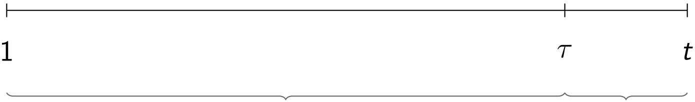
$C_{k,t}$ = cost of optimal segmentation in k segments up to point t.

$1 \qquad\qquad\qquad\qquad\qquad\qquad \tau \qquad\qquad t$

$$\underbrace{\min_\theta \left\{ \ell \left(Y^{[\tau+1,t]} \middle| \theta \right) \right\}}_{\text{traditional likelihood inference}}$$

How it works (with equations)

$C_{k,t}$ = cost of optimal segmentation in k segments up to point t.

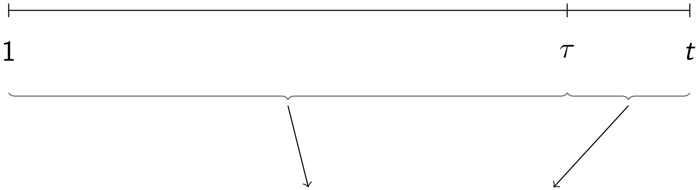


$$C_{k-1,\tau} + \underbrace{\min_\theta \left\{ \ell \left(Y^{[\tau+1,t]} \middle| \theta \right) \right\}}_{\text{traditional likelihood inference}}$$

(Using summation assumption)

How it works (with equations)

$C_{k,t}$ = cost of optimal segmentation in k segments up to point t.



$$C_{k,t} = \mathrm{min}_{\{k-1<\tau<t\}} \left\{ C_{k-1,\tau} + \underbrace{\mathrm{min}_\theta \left\{ \ell\left(Y^{[\tau+1,t]} \middle| \theta \right) \right\}}_{\text{traditional likelihood inference}} \right\}$$

(Using summation assumption)

Choice of K ?

- Standard criteria (AIC, BIC...)
  $\rightarrow$ theoretically unjustified
  $\rightarrow$ asymptotic criteria
  $\rightarrow$ overestimation

Choice of K ?

- Standard criteria (AIC, BIC...)
  $\rightarrow$ theoretically unjustified
  $\rightarrow$ asymptotic criteria
  $\rightarrow$ overestimation

- Oracle inequality on the risk of the models
  $\rightarrow$ Collection of models $\mathcal{S}_m$
  $\rightarrow$ Minimal contrast estimator $\hat{s}_m = \arg\min_{s \in \mathcal{S}_m} \gamma(u)$

$$R(s, \hat{s}_{m(s)})$$

Choice of K ?

- Standard criteria (AIC, BIC...)
  $\rightarrow$ theoretically unjustified
  $\rightarrow$ asymptotic criteria
  $\rightarrow$ overestimation

- Oracle inequality on the risk of the models
  $\rightarrow$ Collection of models $\mathcal{S}_m$
  $\rightarrow$ Minimal contrast estimator $\hat{s}_m = \arg\min_{s \in \mathcal{S}_m} \gamma(u)$
  $\rightarrow$ Penalized-likelihood estimator
  $\hat{m} = \arg\min_{m \in \mathcal{M}_n} \{\gamma(\hat{s}_m) + pen(m)\}$

$$R(s, \hat{s}_{\hat{m}}) \leq C\, R(s, \hat{s}_{m(s)})$$

Specifications

- true model

$$s(t) = \mathcal{NB}(\theta_t, \phi)$$

- collection of models

$$\mathcal{S}_m = \{s_m \mid \forall J \in m,\ \forall t \in J,\ s_m(t) = \mathcal{NB}(\theta_J, \phi)\}$$

- log-likelihood contrast

$$\gamma(s) = -\sum_{t=1}^{n} \phi \log \theta_t + Y_t \log(1 - \theta_t)$$

- Kullback-Leibler risk

Main Theorem

Let $\mathcal{M}_n$ be a collection of partitions constructed on a partition m_f such that there exist absolute positive constants ρ_{min}, ρ_{max} and Γ satisfying:

- $\forall t, \rho_{min} \leq \theta_t \leq \rho_{max}$ and
- $\forall J \in m_f, |J| \geq \Gamma\ (\log n)^2$.

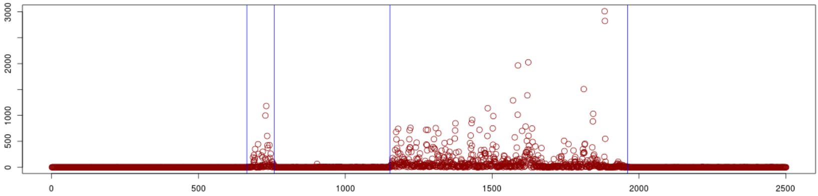
Let $\beta > 1/2\rho_{min}$. If, $\forall m \in \mathcal{M}_n$,

$$pen(m) \geq \beta |m| \left(1 + 4\sqrt{1.1 + \log\left(\frac{n}{|m|}\right)}\right)^2,$$

$$\mathbf{E}\left[h^2(s, \hat{s}_{\hat{m}})\right] \leq C \log(n) \inf_{m \in \mathcal{M}_n} \{\mathbf{E}[K(s, \hat{s}_m)]\} + C(\phi, \Gamma, \rho_{min}, \rho_{max}, \beta, \Sigma).$$

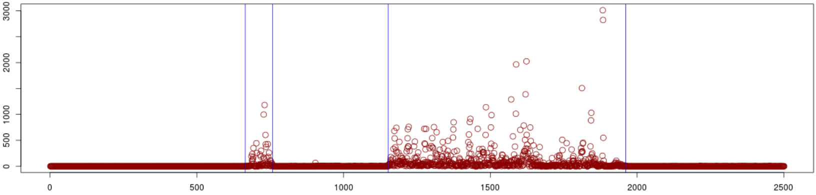
Segmentation procedure

- $K = 5$

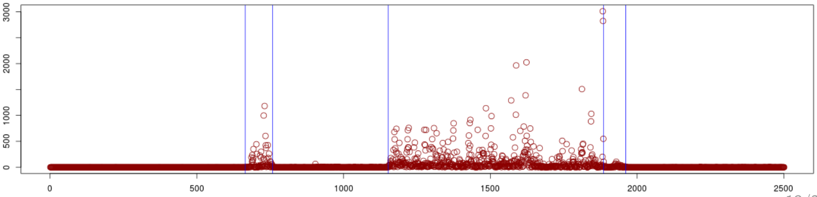


Segmentation procedure

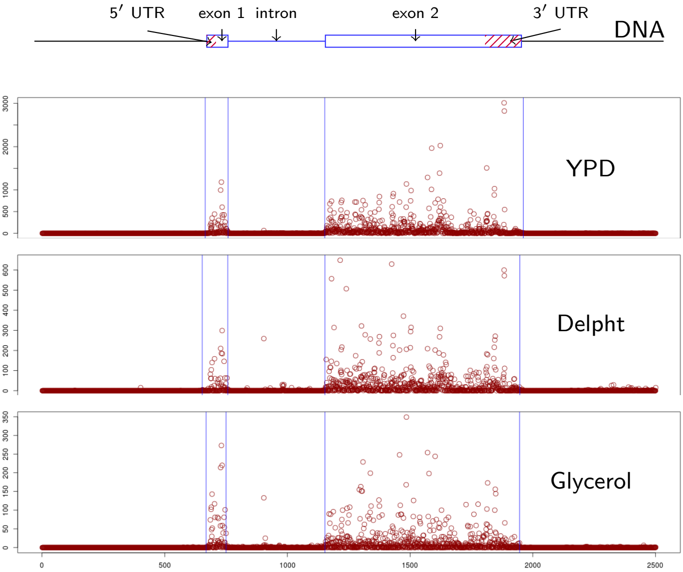
- $K = 5$



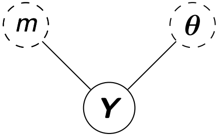
- estimated K $(= 6)$



How to compare different conditions ?



Confidence in the segmentation

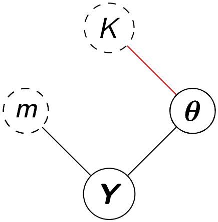


- $\forall J \in m, \forall t \in J, Y_t \sim \mathcal{NB}(\theta_J, \phi)$

Confidence in the segmentation

Rigaill *et.al.* (2012)

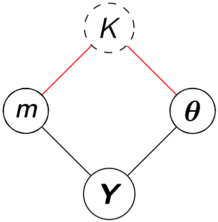
- $\forall i \in \{1, \ldots, K\}$,
  $\theta_i \sim \mathcal{B}eta(1/2, 1/2)$
- $\forall J \in m, \forall t \in J, Y_t \sim \mathcal{NB}(\theta_J, \phi)$



Confidence in the segmentation

Rigaill *et.al.* (2012)

- $m \sim \mathcal{U}\left(\mathcal{M}_K\right)$
- $\forall i \in \{1, \ldots, K\}$,
  $\theta_i \sim \mathcal{B}eta(1/2, 1/2)$
- $\forall J \in m, \forall t \in J, Y_t \sim \mathcal{NB}(\theta_J, \phi)$



Computations

$$\mathbb{P}(Y|K) = \sum_{m \in \mathcal{M}_K} \mathbb{P}(Y, m|K) = \sum_{m \in \mathcal{M}_K} C_K \prod_{J \in m} a_J P(Y_J|J)$$

Computations

$$\mathbb{P}(Y|K) = \sum_{m\in\mathcal{M}_K} \mathbb{P}(Y,m|K) = \sum_{m\in\mathcal{M}_K} C_K \prod_{J\in m} a_J P(Y_J|J)$$

Key point: $\forall 1 \leq i < j \leq n+1$,

$$[A]_{i,j} = a_{[\![i,j[\![}P(Y_{[\![i,j[\![}|[\![i,j[\![) = a_{[\![i,j[\![} \int \mathbb{P}(Y_{[\![i,j[\![}|\theta_{[\![i,j[\![})\mathbb{P}(\theta_{[\![i,j[\![})d\theta_{[\![i,j[\![}$$

- $\mathbb{P}(Y|K) = C_K\ [A^K]_{1,n+1}$;

Computations

$$\mathbb{P}(Y|K) = \sum_{m \in \mathcal{M}_K} \mathbb{P}(Y, m|K) = \sum_{m \in \mathcal{M}_K} C_K \prod_{J \in m} a_J P(Y_J|J)$$

Key point: $\forall 1 \leq i < j \leq n+1$,

$$[A]_{i,j} = a_{[\![i,j[\![} P(Y_{[\![i,j[\![} | [\![i,j[\![) = a_{[\![i,j[\![} \int \mathbb{P}(Y_{[\![i,j[\![} | \theta_{[\![i,j[\![}) \mathbb{P}(\theta_{[\![i,j[\![}) d\theta_{[\![i,j[\![}$$

- $\mathbb{P}(Y|K) = C_K \; [A^K]_{1,n+1}$;
- $P(\tau_k = t|Y, K) = \dfrac{\left[(A)^k\right]_{1,t} \left[(A)^{K-k}\right]_{t,n+1}}{[(A)^K]_{1,n+1}}$.

Computations

$$\mathbb{P}(Y|K) = \sum_{m \in \mathcal{M}_K} \mathbb{P}(Y, m|K) = \sum_{m \in \mathcal{M}_K} C_K \prod_{J \in m} a_J P(Y_J|J)$$

Key point: $\forall 1 \leq i < j \leq n+1$,

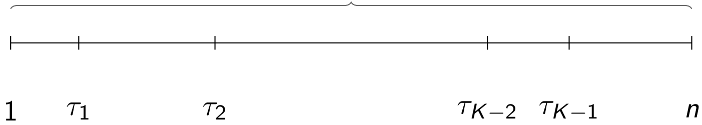
$$[A]_{i,j} = a_{[\![i,j[\![} P(Y_{[\![i,j[\![} | [\![i,j[\![) = a_{[\![i,j[\![} \int \mathbb{P}(Y_{[\![i,j[\![} | \theta_{[\![i,j[\![}) \mathbb{P}(\theta_{[\![i,j[\![}) d\theta_{[\![i,j[\![}$$

- $\mathbb{P}(Y|K) = C_K \; [A^K]_{1,n+1}$;
- $P(\tau_k = t | Y, K) = \dfrac{\left[(A)^k\right]_{1,t} \left[(A)^{K-k}\right]_{t,n+1}}{\left[(A)^K\right]_{1,n+1}}$.

$\mathcal{NB}$ with conjugate prior $\Rightarrow$ exact computations; quadratic time

How it works

All segmentations of $\{1, n\}$ in K segments
with last changepoint τ_{K-1}



How it works

$$\overbrace{\underbrace{1 \qquad \tau_1 \qquad \tau_2 \qquad \qquad \tau_{K-2} \quad \tau_{K-1}}_{\text{All segmentations of } \{1,\tau_{K-1}\} \text{ in } K-1 \text{ segments}} \underbrace{\qquad \qquad n}_{\text{last segment}}}^{\text{All segmentations of } \{1,n\} \text{ in } K \text{ segments with last changepoint } \tau_{K-1}}$$

How it works (with equations)

Let $P(k)_{1,t}$ = posterior probability of k segments up to point t.

$1 \qquad \tau \qquad t$

$$\underbrace{P\left(Y^{[\tau+1,t]} \middle| [\tau+1,t] \right)}_{= \int_\theta P\left(Y^{[\tau+1,t]} \middle| \theta \right) dp(\theta)}$$

How it works (with equations)

Let $P(k)_{1,t}$ = posterior probability of k segments up to point t.

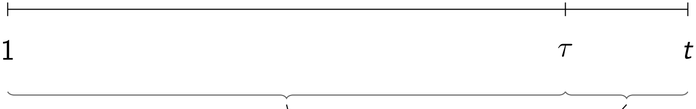
$1 \qquad \tau \qquad t$

$$P(k-1)_{1,\tau} \times \underbrace{P\left(Y^{[\tau+1,t]} \middle| [\tau+1,t] \right)}_{= \int_\theta P\left(Y^{[\tau+1,t]} \middle| \theta \right) dp(\theta)}$$

(Using factorability assumption)

How it works (with equations)

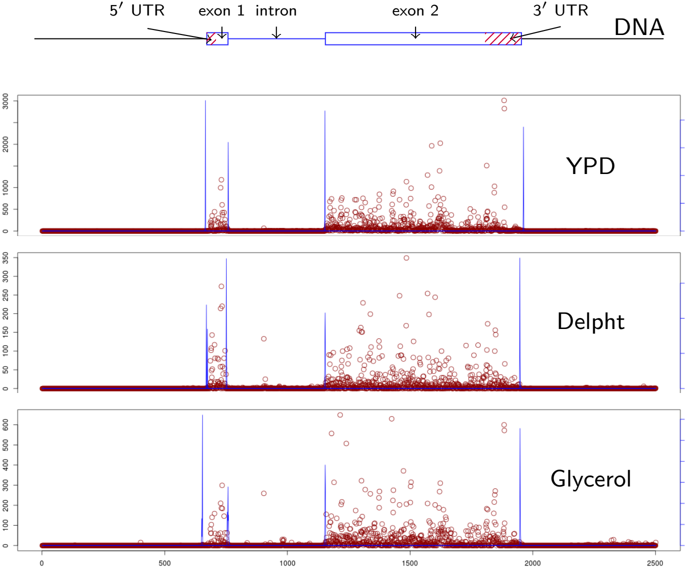
Let $P(k)_{1,t}$ = posterior probability of k segments up to point t.



$$P(k)_{1,t} = \sum_{\{k-1<\tau<t\}} \left\{ P(k-1)_{1,\tau} \times \underbrace{P\left(Y^{[\tau+1,t]} \middle| [\tau+1,t] \right)}_{= \int_\theta P\left(Y^{[\tau+1,t]} \middle| \theta \right) dp(\theta)} \right\}$$

(Using factorability assumption)

How to compare different conditions ?



Model for two profiles

- Series Y^1 and Y^2 with same length n
- Respective (known) number of segments K^1 and K^2.
- Change-points to compare: $\tau^1_{k_1}$ and $\tau^2_{k_2}$

Model for two profiles

- Series Y^1 and Y^2 with same length n
- Respective (known) number of segments K^1 and K^2.
- Change-points to compare: $\tau^1_{k_1}$ and $\tau^2_{k_2}$

$\rightarrow \Delta = \tau^1_{k_1} - \tau^2_{k_2}$

$$P(\Delta = d | Y^1, Y^2, K^1, K^2) = \sum_t P(\tau_{k_1} = t | Y^1, K^1) P(\tau_{k_2} = t - d | Y^2, K^2)$$

Still exact computations, quadratic time

Position of 0 w.r.t. posterior distribution of shift?

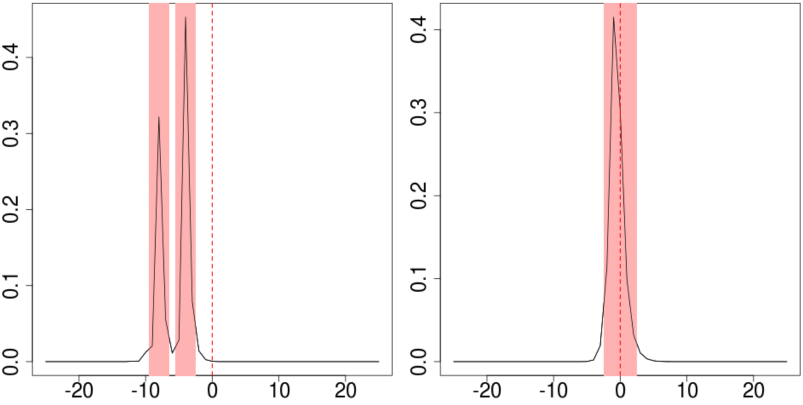


Figure: posterior distribution and 95% credibility intervals of shift

Framework of I profiles

- Series Y^ℓ for $1 \leq \ell \leq I$ $\rightarrow \mathbf{Y}$
- Partitions m^ℓ in K^ℓ segments $\rightarrow \mathbf{m}$ and $\mathbf{K}$
- Parameters θ^ℓ $\rightarrow \boldsymbol{\theta}$
- Change-points to compare: $\tau^\ell_{k_\ell}$

Framework of I profiles

- Series Y^ℓ for $1 \leq \ell \leq I$ $\rightarrow$ $\mathbf{Y}$
- Partitions m^ℓ in K^ℓ segments $\rightarrow$ $\mathbf{m}$ and $\mathbf{K}$
- Parameters θ^ℓ $\rightarrow$ $\boldsymbol{\theta}$
- Change-points to compare: $\tau^\ell_{k_\ell}$

- $E_0 = \{\tau^1_{k_1} = \cdots = \tau^I_{k_I}\}$.
- $\mathbf{E} = \mathbb{I}\{E_1\} = 1 - \mathbb{I}\{E_0\}$.

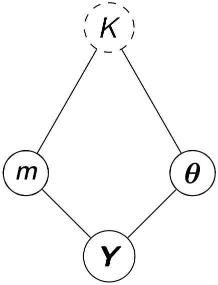
Under the product model,

$$P(\mathbf{Y}|\mathbf{K}) = \prod_\ell P(Y^\ell|K^\ell)$$

$$P(E_0|\mathbf{K}) \simeq 10^{-8}$$

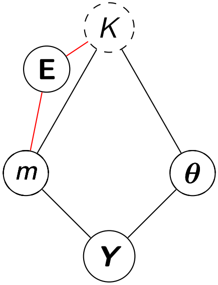
Confidence in the segmentation

- $m \sim \mathcal{U}(\mathcal{M}_K)$
- $\forall i \in \{1, \ldots, K\}, \theta_i \sim \mathcal{B}eta(1/2, 1/2)$
- $\forall J \in m, \forall t \in J, Y_t \sim \mathcal{NB}(\theta_J, \phi)$



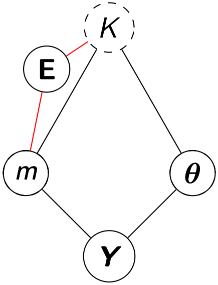
Confidence in the segmentation

- $\mathbf{E} \sim \mathcal{B}(p_0)$
- $\mathbf{m}|\mathbf{K},\mathbf{E} \sim P(\mathbf{m}|\mathbf{K},\mathbf{E})$
- $\forall i \in \{1,\ldots,K\}, \theta_i \sim \mathcal{B}eta(1/2, 1/2)$
- $\forall J \in m, \forall t \in J, Y_t \sim \mathcal{NB}(\theta_J, \phi)$



Confidence in the segmentation

- $\mathbf{E} \sim \mathcal{B}(p_0)$
- $\mathbf{m}|\mathbf{K}, \mathbf{E} \ \sim P\left(\mathbf{m}|\mathbf{K}, \mathbf{E}\right)$
- $\forall i \in \{1, \ldots, K\}, \theta_i \sim \mathcal{B}eta(1/2, 1/2)$
- $\forall J \in m, \forall t \in J, Y_t \sim \mathcal{NB}(\theta_J, \phi)$



For instance,

- $\mathbf{m}|\mathbf{K}, E_0 = \mathcal{U}\left(\mathcal{M}_{n,\mathbf{K}} \cap E_0\right)$
- $\mathbf{m}|\mathbf{K}, E_1 = \mathcal{U}\left(\mathcal{M}_{n,\mathbf{K}} \cap E_1\right)$

where $\mathcal{M}_{n,\mathbf{K}} = \bigotimes_\ell \mathcal{M}_{n,K^\ell}$.

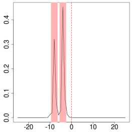
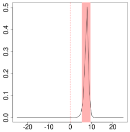
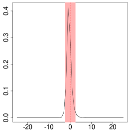
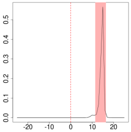
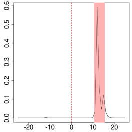
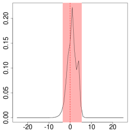
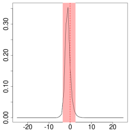
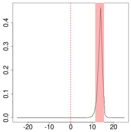
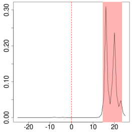
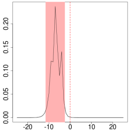
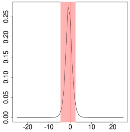
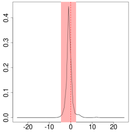
Posterior probability of E_0

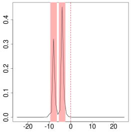
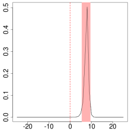
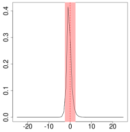
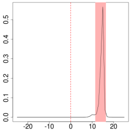
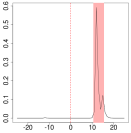
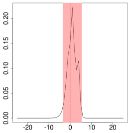
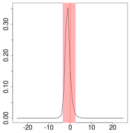
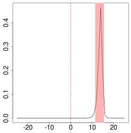
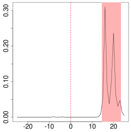
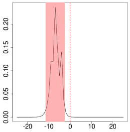
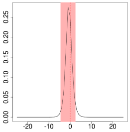
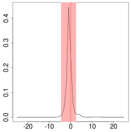
$$P(E_0|\mathbf{Y},\mathbf{K}) = \frac{p_0(\mathbf{K})}{q_0(\mathbf{K})} Q(\mathbf{Y},E_0|\mathbf{K}) \Bigg/ \left[\frac{p_0(\mathbf{K})}{q_0(\mathbf{K})} Q(\mathbf{Y},E_0|\mathbf{K}) + \frac{1-p_0(\mathbf{K})}{1-q_0(\mathbf{K})} Q(\mathbf{Y},E_1|\mathbf{K}) \right]$$

- $Q(\mathbf{Y},E_0|\mathbf{K})$ and $Q(\mathbf{Y},E_1|\mathbf{K})$
  $\rightarrow$ probabilites induced by product model
  $\rightarrow$ computed exactly with probability matrices A_ℓ
- $\frac{p_0(\mathbf{K})}{q_0(\mathbf{K})}, \frac{1-p_0(\mathbf{K})}{1-q_0(\mathbf{K})}$
  $\rightarrow$ weights induced by new model
  $\rightarrow$ $q_0(\mathbf{K})$ prior under product model; $p_0(\mathbf{K})$ under new model

$\rightarrow$ exact computation, and in quadratic time

$\rightarrow$ R package `EBS`

	5' UTR	5' intron	3' intron	3' UTR
ypd/del				
ypd/gly				
del/gly				

	5' UTR	5' intron	3' intron	3' UTR
ypd/del				
ypd/gly				
del/gly				
$P(E_0 \mid \mathbf{Y}, \mathbf{K})$	10^{-3}	0.99	0.99	$6\ .10^{-3}$

$$P(E_0 \mid \mathbf{K}) = 1/2$$

Illustration: 50 genes with 2 exons

Distribution of the posterior probability of event E_0

$a) \quad p_0(K) = 1/2$

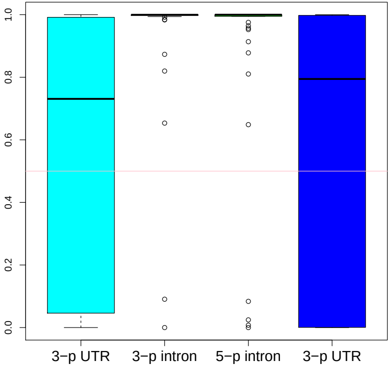
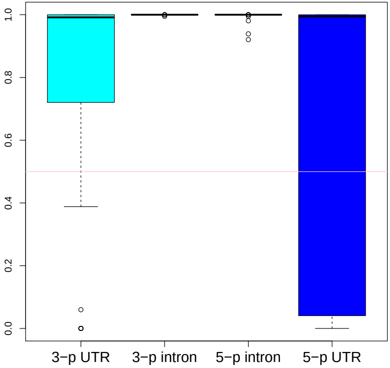


Illustration: 50 genes with 2 exons

Distribution of the posterior probability of event E_0

$b)$ $p_0(K) = 0.9$ for UTRs, $p_0(K) = 0.99$ for introns



Acknowledgements

- Stéphane Robin (INRA, AgroParisTech)
- Émilie Lebarbier (AgroParisTech)
- Guillem Rigaill (Université d'Evry)
- Gavin Sherlock (Standford University)

Thank you !