Two segmentation methods for genome annotation

Alice Cleynen

UMR 518 AgroParisTech / UCBerkeley Biostatistics
Stéphane Robin / Sandrine Dudoit

June 12th, 2013

RNA-Seq data and genome annotation

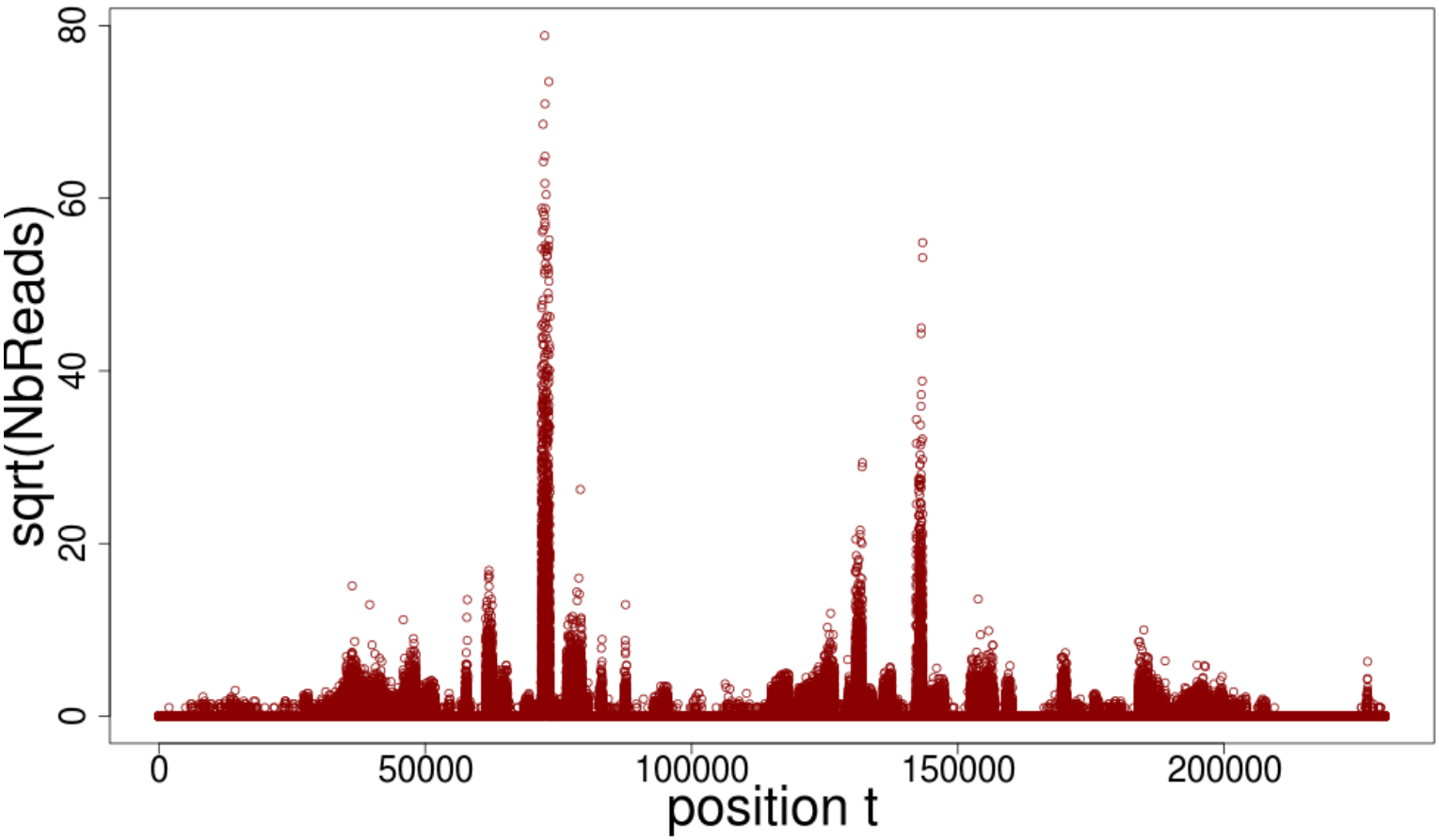
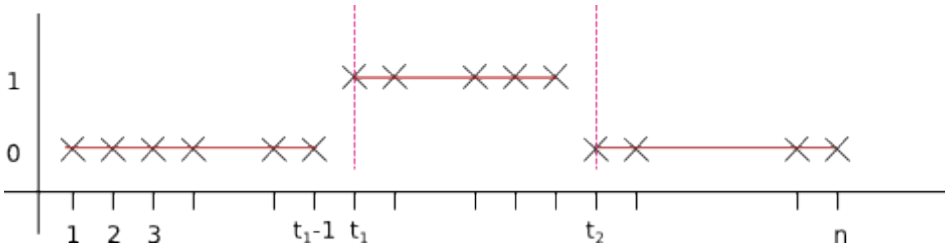


Figure: Number of reads starting at position t

Outline

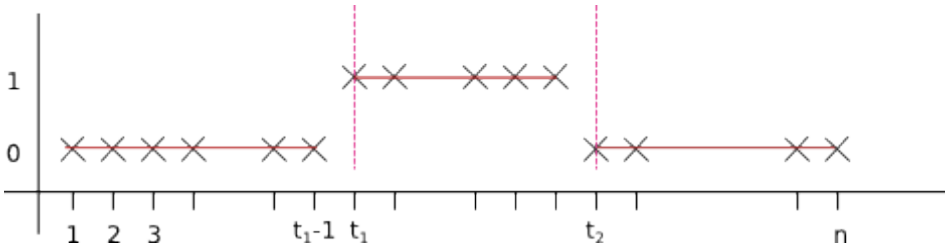
Segmentation context



- m a partition of $[\![1, n]\!]$
- J a segment of m

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

Segmentation context

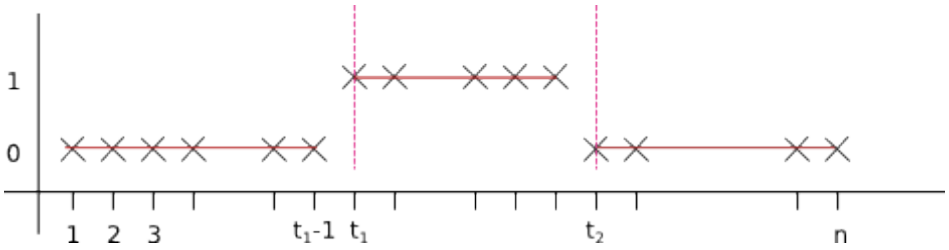


- m a partition of $[\![1, n]\!]$
- J a segment of m

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

$\rightarrow$ Can we find the optimal segmentation in K segments?
(with respect to the log-likelihood criterion)

Segmentation context



- m a partition of $[\![1, n]\!]$
- J a segment of m

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

$\rightarrow$ Can we find the optimal segmentation in K segments?
(with respect to the log-likelihood criterion)

$\rightarrow$ How can we choose K?

Finding the optimal segmentation

Issues:

- size of the data-set quite large (10^6 to 10^8)
  $\rightarrow$ prohibits large complexity algorithms

Finding the optimal segmentation

Issues:

- size of the data-set quite large (10^6 to 10^8)
  $\rightarrow$ prohibits large complexity algorithms
- change-points are discrete parameters
  $\rightarrow$ exploration of the whole segmentation space. $\binom{n-1}{K-1}$ partitions

Finding the optimal segmentation

Issues:

- size of the data-set quite large (10^6 to 10^8)
  $\rightarrow$ prohibits large complexity algorithms
- change-points are discrete parameters
  $\rightarrow$ exploration of the whole segmentation space. $\binom{n-1}{K-1}$ partitions
- count dataset
  $\rightarrow$ few available algorithms

Finding the optimal segmentation

Issues:

- size of the data-set quite large (10^6 to 10^8)
  $\rightarrow$ prohibits large complexity algorithms
- change-points are discrete parameters
  $\rightarrow$ exploration of the whole segmentation space. $\binom{n-1}{K-1}$ partitions
- count dataset
  $\rightarrow$ few available algorithms

pruned DPA
(Rigaill [5])

- fast (empirically linear)
- exact (maximizes the log-likelihood)

Finding the optimal segmentation

Issues:

- size of the data-set quite large (10^6 to 10^8)
  $\rightarrow$ prohibits large complexity algorithms
- change-points are discrete parameters
  $\rightarrow$ exploration of the whole segmentation space. $\binom{n-1}{K-1}$ partitions
- count dataset
  $\rightarrow$ few available algorithms

pruned DPA
(Rigaill [5])

- fast (empirically linear)
- exact (maximizes the log-likelihood)

Valid for one-parameter losses from exponential family

$$\mathcal{NB}(p_J, \phi)$$

$\rightarrow$ Need to estimate overdispersion parameter ϕ.

Estimation of overdispersion parameter

Inspired from Johnson, Kotz and Kemp [3]

- Let h equal 15;
- for each sliding window L of size h, compute moment estimator of ϕ_L using

$$\phi_L = \frac{\mathbf{E}^2(X^L)}{\mathbf{V}(X^L) - \mathbf{E}(X^L)}$$

- if $\hat{\phi} = median\{\phi_L\} < 0$, $h \leftarrow h * 2$, else keep $\hat{\phi}$

Estimation of overdispersion parameter

Inspired from Johnson, Kotz and Kemp [3]

- Let h equal 15;
- for each sliding window L of size h, compute moment estimator of ϕ_L using

$$\phi_L = \frac{\mathbf{E}^2(X^L)}{\mathbf{V}(X^L) - \mathbf{E}(X^L)}$$

- if $\hat{\phi} = median\{\phi_L\} < 0$, $h \leftarrow h * 2$, else keep $\hat{\phi}$

$\rightarrow$ good results in practice
$\rightarrow$ no theoretical guarantees
$\rightarrow$ ongoing work

Choosing the number of segments

joint work with E. Lebarbier

$\rightarrow$ penalized likelihood framework (Inspired by Birgé and Massart [2]):

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

$$\hat{m} = \arg\min_{m \in \mathcal{M}_n} \{-\text{log-lik}(\hat{s}_m) + pen(m)\},$$

Choosing the number of segments

joint work with E. Lebarbier

$\rightarrow$ penalized likelihood framework (Inspired by Birgé and Massart [2]):

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

$$\hat{m} = \arg \min_{m \in \mathcal{M}_n} \{-\text{log-lik}(\hat{s}_m) + pen(m)\},$$

$$\mathbb{E}\left[h^2(s, \hat{s}_{\hat{m}})\right] \leq C \log(n) \inf_{m \in \mathcal{M}_n} \{\mathbb{E}[K(s, \hat{s}_m)]\} + Cst.$$

Proposed Procedure

R package `Segmentor3IsBack`

1. estimate ϕ
2. choose K_{max}
3. for $1 \leq k \leq K_{max}$, compute $\hat{s}_k$ (with pDPA)
4. choose $\hat{K} = \arg\min_k \left\{ -\text{log-lik}(\hat{s}_k) + \beta k \left(1 + 4\sqrt{1.1 + \log \frac{n}{k}}\right)^2 \right\}$
   (tune β using the slope heuristic[1])

Yeast analysis

$\hat{K} = 103$ segments

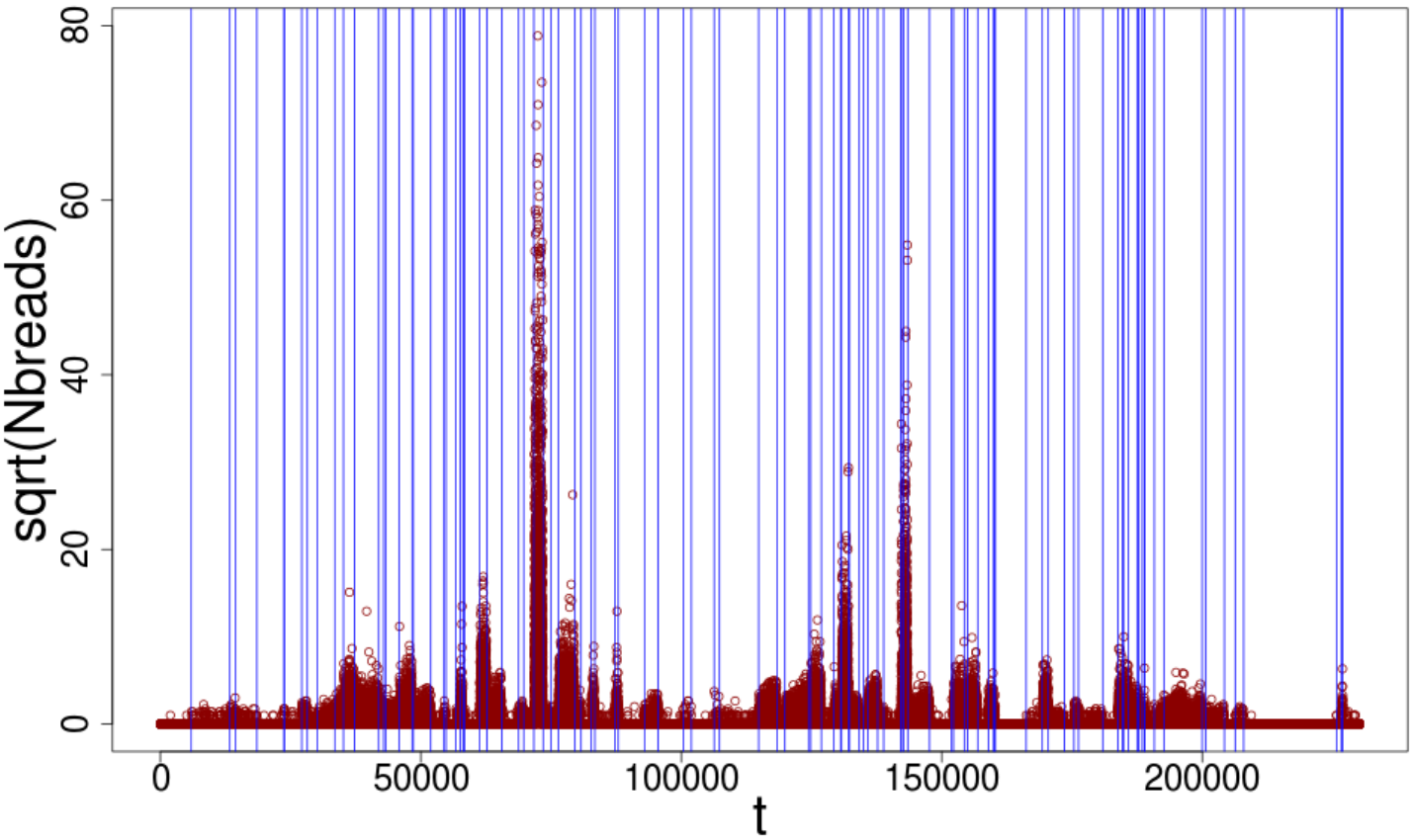


Figure: Segmentation of S. Cerevisiae chromosome 1 (positive strand)

Outline

A need for gene re-annotation

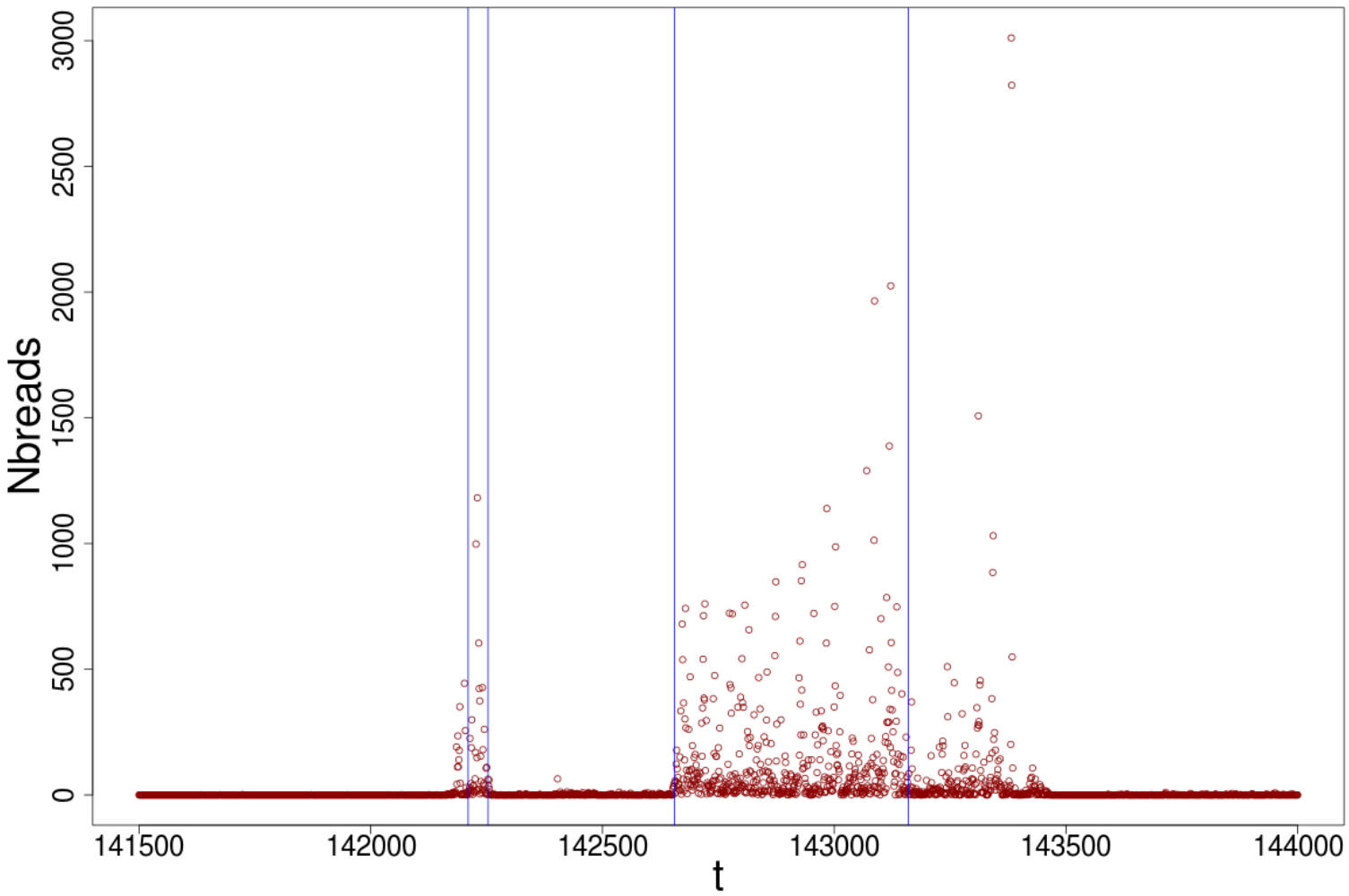


Figure: Yeast EFB1 gene and SGD official annotation

Framework

Bayesian segmentation approach introduced by Rigaill *et al.* [4]

$$\begin{aligned} Y_t &\sim \mathcal{NB}(p_J, \phi) \quad \text{if } t \in J \\ p_J &\sim \mathcal{B}eta(a, b) \end{aligned}$$

- Suppose counts independent conditionally to the parameters.
- Suppose K is known; $\tau_k =$ k^{th} breakpoint.
- if $J = [\![t_1, t_2 [\![$, let $Y_J = [\![Y_{t_1}, Y_{t_2} [\![$.

Posterior Probabilities

$$\mathbb{P}(m, Y) = \mathbb{P}(K)\mathbb{P}(m|K)\mathbb{P}(\phi|m) \prod_{J \in m} \int \mathbb{P}(Y_J|p_J, \phi)\mathbb{P}(p_J)dp_J$$

Posterior Probabilities

$$\mathbb{P}(m, Y) = \mathbb{P}(K)\mathbb{P}(m|K)\mathbb{P}(\phi|m) \prod_{J \in m} \int \mathbb{P}(Y_J|p_J, \phi)\mathbb{P}(p_J)dp_J$$

- requires knowledge of ϕ: use estimator defined above

Posterior Probabilities

$$\mathbb{P}(m, Y) = \mathbb{P}(K)\mathbb{P}(m|K)\mathbb{P}(\phi|m) \prod_{J \in m} \int \mathbb{P}(Y_J|p_J, \phi)\mathbb{P}(p_J)dp_J$$

- requires knowledge of ϕ: use estimator defined above
- computed exactly, stored in matrix $[A]_{i,j} = [\mathbb{P}(Y_{[\![i,j]\!]}|[\![i,j[\![, \phi)]$

Posterior Probabilities

$$\mathbb{P}(m, Y) = \mathbb{P}(K)\mathbb{P}(m|K)\mathbb{P}(\phi|m) \prod_{J \in m} \int \mathbb{P}(Y_J|p_J, \phi)\mathbb{P}(p_J)dp_J$$

- requires knowledge of ϕ: use estimator defined above
- computed exactly, stored in matrix $[A]_{i,j} = [\mathbb{P}(Y_{[\![i,j[\![}|[\![i,j[\![, \phi)]$

$\rightarrow$ compute $\mathbb{P}(Y|K)$ as $C\ [A^K]_{1,n+1}$

Posterior Probabilities

$$\mathbb{P}(m, Y) = \mathbb{P}(K)\mathbb{P}(m|K)\mathbb{P}(\phi|m) \prod_{J \in m} \int \mathbb{P}(Y_J|p_J, \phi)\mathbb{P}(p_J)dp_J$$

- requires knowledge of ϕ: use estimator defined above
- computed exactly, stored in matrix $[A]_{i,j} = [\mathbb{P}(Y_{[\![i,j]\!]}|[\![i, j[\![, \phi)]$

$\rightarrow$ compute $\mathbb{P}(Y|K)$ as $C\ [A^K]_{1,n+1}$
$\rightarrow$ compute, for all $1 \leq t \leq n$, $p(\tau_k = t|Y, K)$ and the associated 95% credibility intervals.

Posterior Probabilities

$$\mathbb{P}(m, Y) = \mathbb{P}(K)\mathbb{P}(m|K)\mathbb{P}(\phi|m) \prod_{J \in m} \int \mathbb{P}(Y_J|p_J, \phi)\mathbb{P}(p_J)dp_J$$

- requires knowledge of ϕ: use estimator defined above
- computed exactly, stored in matrix $[A]_{i,j} = [\mathbb{P}(Y_{[\![i,j[\![}|[\![i,j[\![, \phi)]$

$\rightarrow$ compute $\mathbb{P}(Y|K)$ as $C\ [A^K]_{1,n+1}$
$\rightarrow$ compute, for all $1 \leq t \leq n$, $p(\tau_k = t|Y, K)$ and the associated 95% credibility intervals.

Issues: numerical precision; quadratic complexity

Posterior Probabilities

$$\mathbb{P}(m, Y) = \mathbb{P}(K)\mathbb{P}(m|K)\mathbb{P}(\phi|m) \prod_{J \in m} \int \mathbb{P}(Y_J|p_J, \phi)\mathbb{P}(p_J)dp_J$$

- requires knowledge of ϕ: use estimator defined above
- computed exactly, stored in matrix $[A]_{i,j} = [\mathbb{P}(Y_{[\![i,j[\![}|[\![i,j[\![, \phi)]$

$\rightarrow$ compute $\mathbb{P}(Y|K)$ as $C\ [A^K]_{1,n+1}$
$\rightarrow$ compute, for all $1 \leq t \leq n$, $p(\tau_k = t|Y, K)$ and the associated 95% credibility intervals.

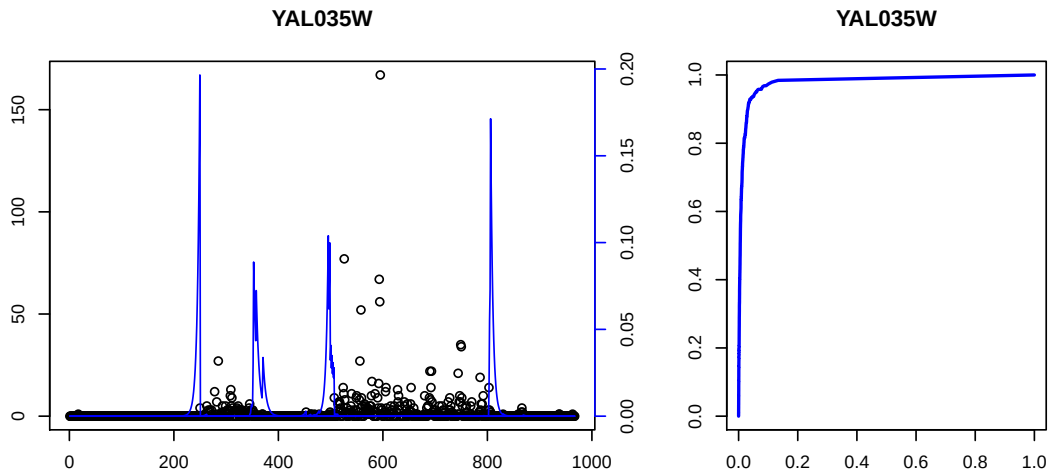
Issues: numerical precision; quadratic complexity

$\rightarrow$ R package `EBS`

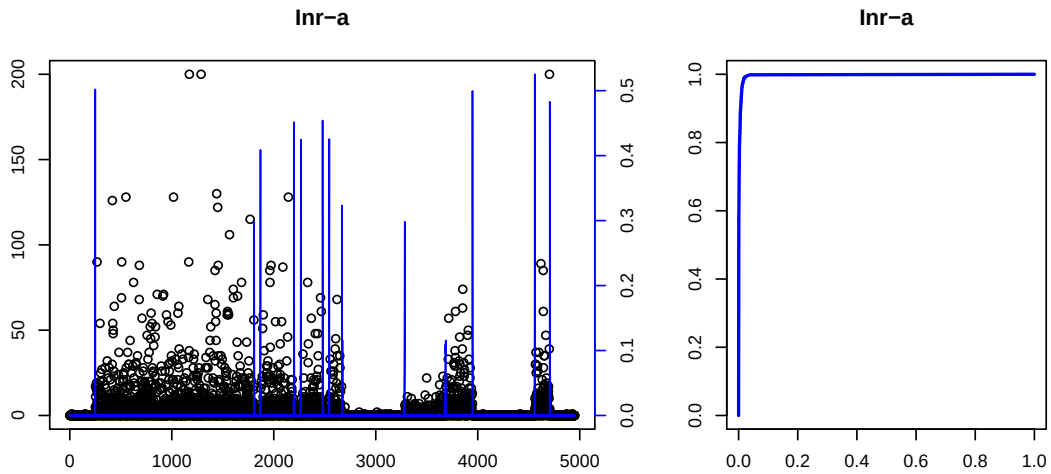
Simulation Study

Two simulation studies:

- Yeast gene with two exons



- Drosophila gene with two isoforms (14 segments total)



Conclusion

Exact segmentation methods for $\mathcal{NB}$ distribution to address

- Whole genome analysis
  - → fast algorithm
  - → choice of number of segments

→ `Segmentor3IsBack`

- Genome re-annotation
  - → confidence in change-point location

→ `EBS`

References

[1] S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.*, 10:245–279 (electronic), 2009.

[2] L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268, 2001.

[3] N. Johnson, A.W. Kemp, and S. Kotz. Univariate discrete distributions. *John Wiley & Sons, Inc.*, 2005.

[4] G. Rigaill, E. Lebarbier, and S. Robin. Exact posterior distributions and model selection criteria for multiple change-point detection problems. *Statistics and Computing*, 22:917–929, 2012.

[5] Guillem Rigaill. Pruned dynamic programming for optimal multiple change-point detection. *Arxiv preprint arXiv:1004.0887*, under review.