A fast algorithm for multiple change-point detection and application to NGS data

Alice Cleynen

January 26th, 2012

Finding Transcripts positions along the genome

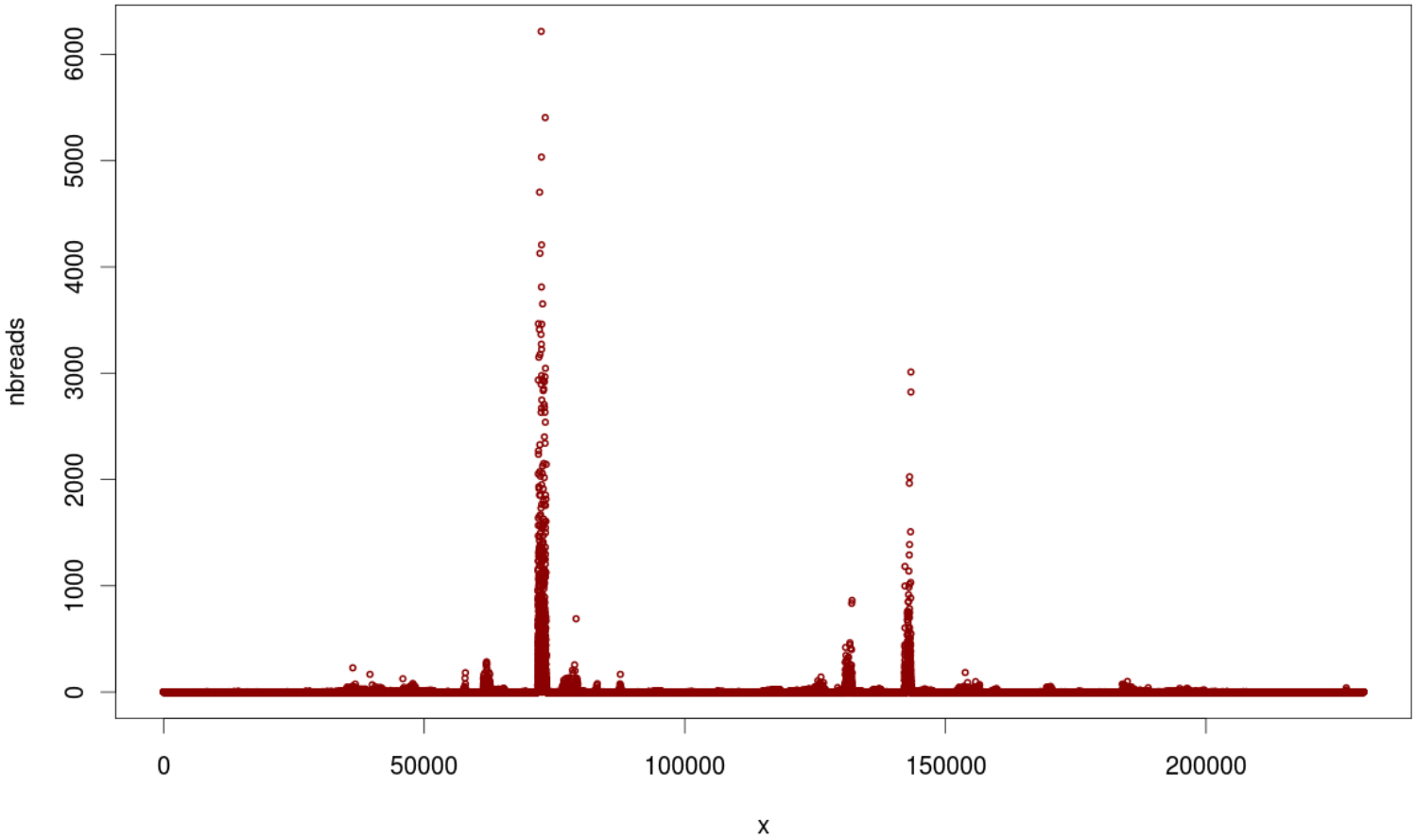
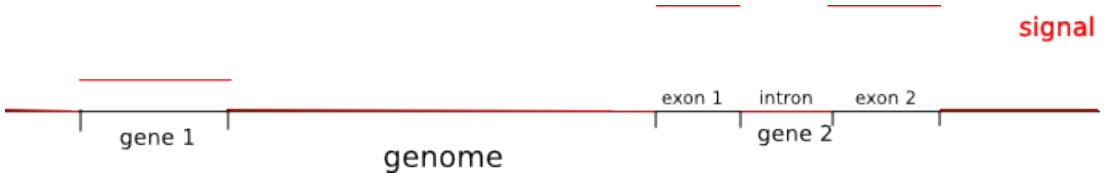


Figure: Number of reads starting at position x

A segmentation context

We assume that the data are a realization of an independant random process which parameters are piece-wise constant.



The intensity of the process depends on the position along the genome (coding or non-coding region) as well as the level of expression of the gene.

References for Negative Binomial model.

A segmentation context

- The signal is divided into K segments $[\tau_{k-1}; \tau_k[, \quad 1 \leq k \leq K$
- The process is drawn from a probability distribution $\mathcal{G}(\theta_r)$
- $\mathcal{G}(\theta_r) \sim \mathcal{BN}(\mu_r, \phi)$

Parameters to estimate:

- Discrete parameters $\{\tau_k, 1 \leq k \leq K\}$
- Continuous parameters $\{\mu_{r_k}, 1 \leq k \leq K\}$
- Continuous parameter ϕ

$$I(K, m, \{\mu_r\}, \phi) = \sum_{r \in m} \sum_{t \in r} \left[-Y_t \log(\mu_r) - \phi \log(1 - \mu_r) - \log \left(\frac{\Gamma(Y_t + \phi)}{\Gamma(\phi) Y_t!} \right) \right]$$

From the original DPA to the Pruned DPA

One-parameter contrast:

$$\gamma(Y_i, \mu) \propto -\log(P(Y_i\ ,\ \mu))$$

Original DPA: segment additivity $\Theta(Kn^2)$

$$C_{k,t} = \min_{\{k-1<\tau<t\}} \left\{ C_{k-1,\tau} + \min_{\mu} \{ \sum_{i=\tau+1}^{t} \gamma(Y_i, \mu) \} \right\}$$

Pruned DPA: point additivity

$$C_{k,t} = \min_{\mu} \left\{ \min_{\{k-1<\tau<t\}} \{ C_{k-1,\tau} + \sum_{i=\tau+1}^{t} \gamma(Y_i, \mu) \} \right\}$$

Two-parameter Negative Binomial Law

- For a given segmentation $\hat{m}$ we have

$$L_{\hat{m}}(\phi) \quad = \quad l(K, \hat{m}, \{\mu_r\}(\phi), \phi)$$

  - Continuous, convex then concave, unique global minimum
  - Easy to optimize using iterative algorithm (eg Armijo's)

- For a given parameter $\hat{\phi}$

$$L_{\hat{\phi}}(m, \mu) \quad = \quad l(K, m, \{\mu_r\}, \hat{\phi})$$

  - 'Easy' to optimize using PDPA algorithm
  - Complexity in $\mathcal{O}(Kn\log(n))$

Global optimization

$$\inf_{m,\{\mu_r\},\phi} l(K, m, \{\mu_r\}, \phi) \quad = \quad \inf_{\phi} \inf_{m \in \mathcal{M}_K^n} \{L_m(\phi)\} \quad = \quad \inf_{\phi} L(\phi)$$

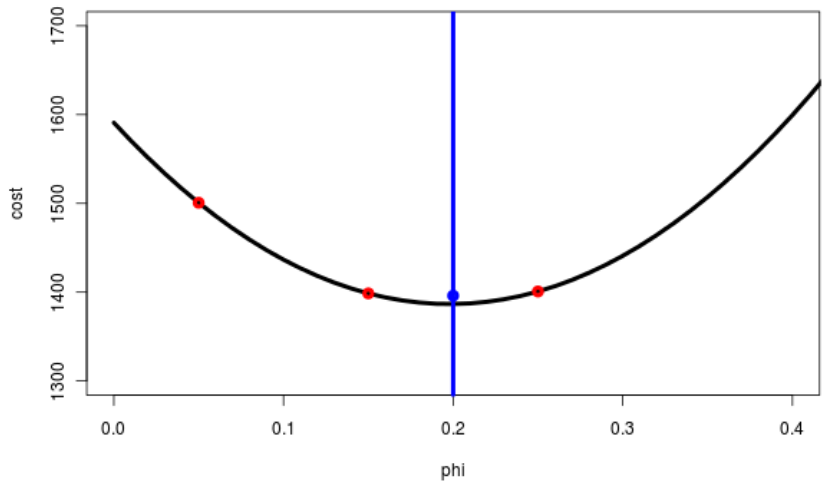
$L(\phi)$ is continuous and has one (or more) global minimum

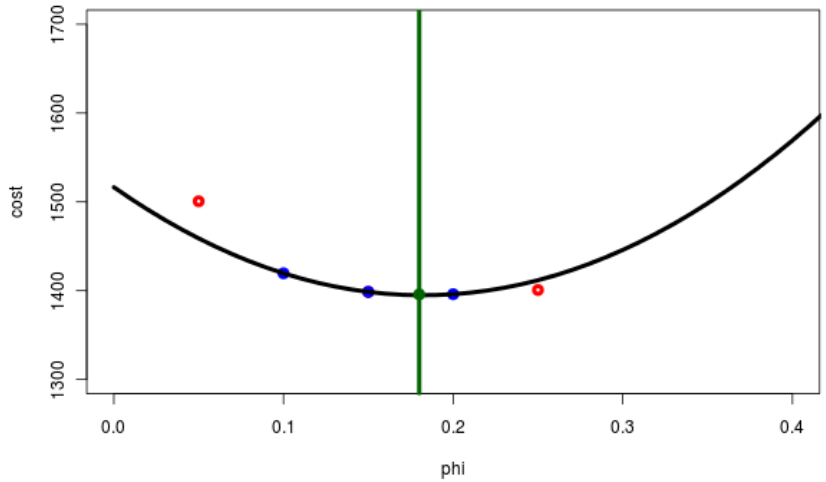
Two heuristic strategies relying on Armijo's algorithm

Strategy 1

Apply Armijo's algorithm directly on $l(K, m, \{\mu_r\}, \phi)$

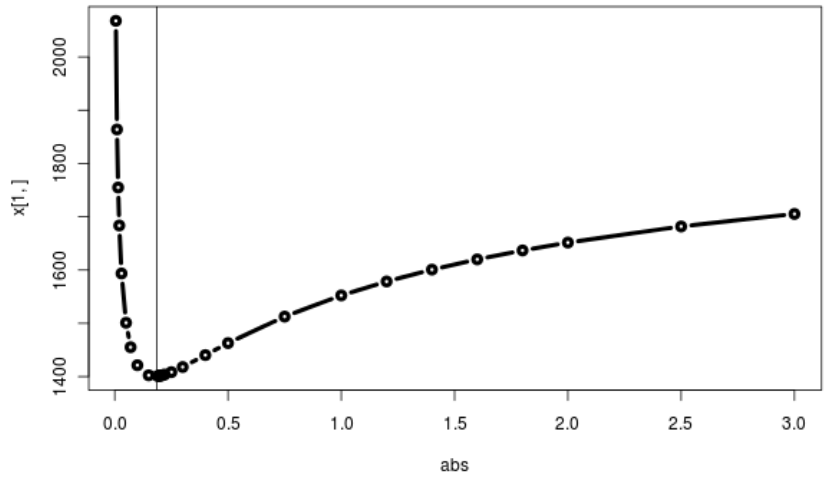
1. Choose three equally spaced ϕ_0, ϕ_1, ϕ_2
2. at step $p+1$
   - run PDPA for each ϕ_i to compute cost
   - fit best quadratic function P
   - find minimum ϕ_{new} of P, run PDPA and find new cost
   - update values of ϕ_0, ϕ_1, ϕ_2
   - if $\phi_{new} = \phi_{old}$, stop

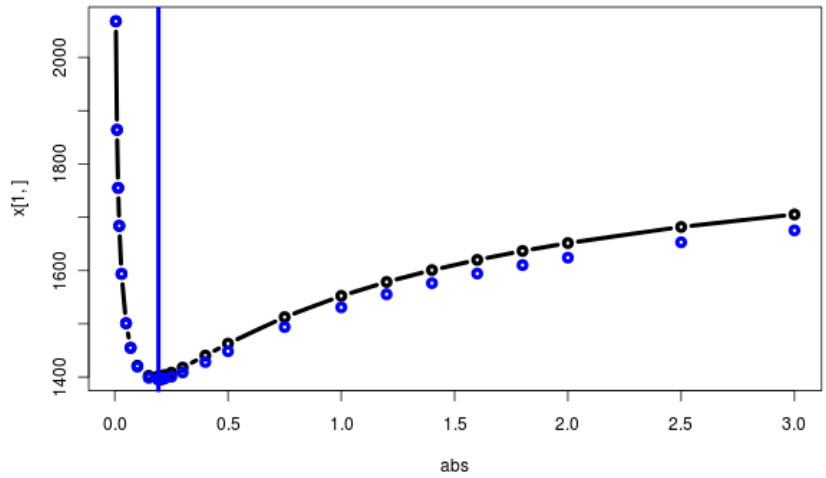




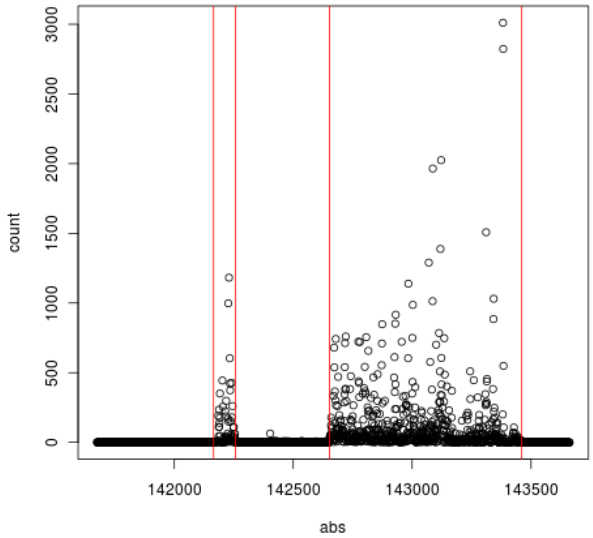
Strategy 2

1. Choose a $\phi_{start} = \phi_0$
2. at step $p+1$
   - $m_p, \mu_p = \text{argmin } L_{\phi_p}(m, \mu)$ using PDPA
   - Update value $\phi_{p+1} = \text{argmin } L_{m_p}(\phi)$ using Armijo's algorithm
   - if $\phi_{p+1} = \phi_p$, stop





Results



Both strategies recover the minimum. BUT:

Strategy 1

- Computationally expensive

Strategy 2

- Might lead to local minimum

Finding Transcripts positions along the genome

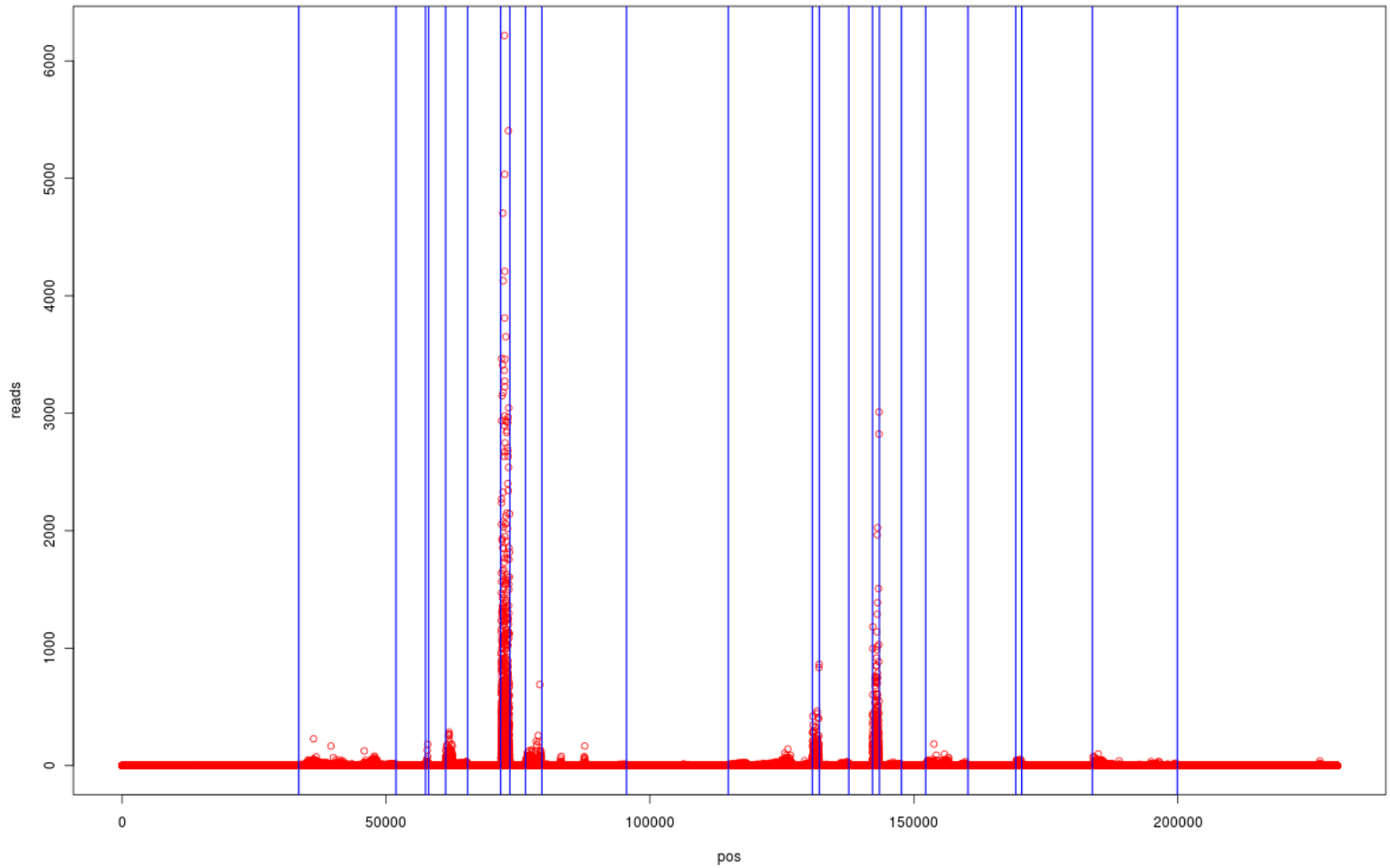


Figure: Segmentation of the chromosome into 25 segments

Conclusion

Open Questions

- Choice of K
- Hypothesis ϕ constant over the signal?

Code:

- R package soon available on the CRAN
- C++ code soon available on the INRA webpage

Thanks!

- Guillem Rigaill
- Michel Koskas
- Stéphane Robin
- You