ENS DE LYON

Identifying a piecewise affine signal from its nonlinear observation - application to DNA replication analysis

Clara Lage[1], Rémi Gribonval[2], Nelly Pustelnik[1], Benjamin Audit[1] and Jean-Michel Arbona[1,3]

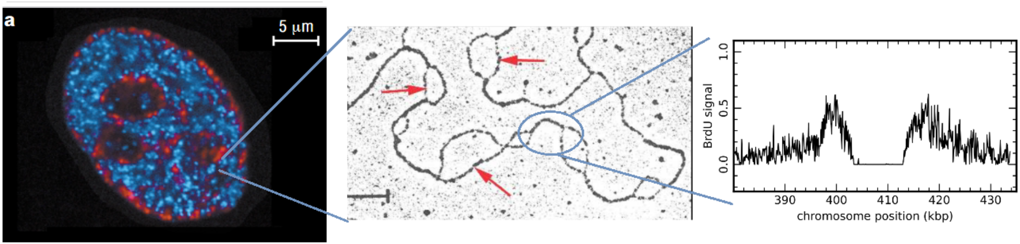
[1] ENS de Lyon, CNRS; [2] LIP, ENS de Lyon; [3] LBMC, Univ de Lyon

Toulouse, October 2024

DNA replication analysis

General framework

- Microscope images are widely used to study the structure and function of cells
- Some characteristics of the replication can be studied by images



late replication
early replication

- This work aims to understand parameters of replication such as: position of origins, local speed, and replication direction

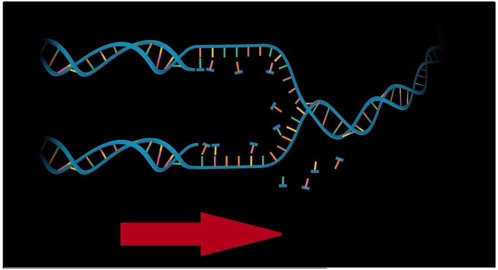
Important parameters

Main Objective: Better characterize the replication of DNA.

- Position of origins of replication
- Speed and direction of replication

Related application:

- Characterization of replication stress in cancer cells



DNA synthesis at a replication fork moving to the right

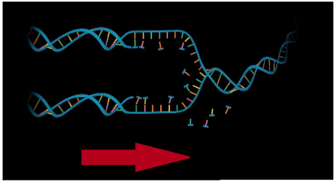
Important parameters

Main Objective: Better characterize the replication of DNA.

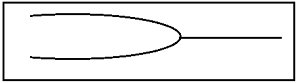
- Position of origins of replication
- Speed and direction of replication

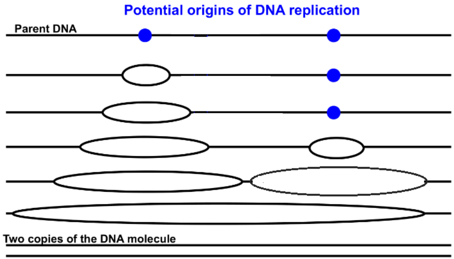
Related application:

- Characterization of replication stress in cancer cells



DNA synthesis at a replication fork moving to the right

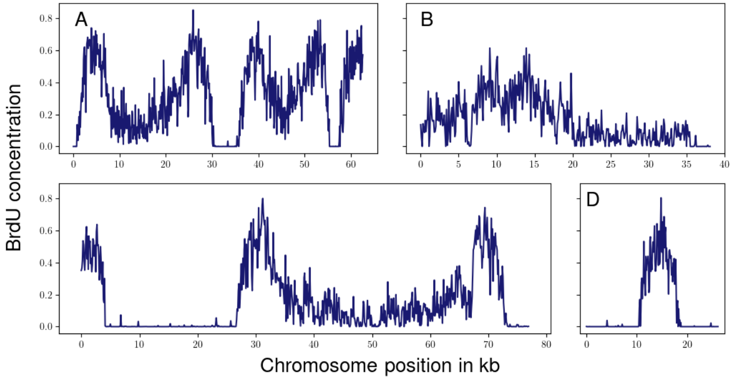




DNA replication signal

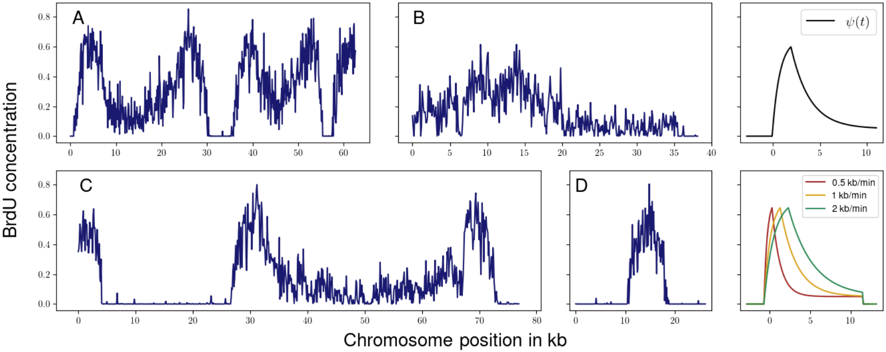
DNA Replication signal

- 1d signals generated during the replication process
- Singals combine DNA sequecing and the concentration of a chemical: BrdU



Dictionary approach

➔ ψ can be used as an atom that can be translated and dilated

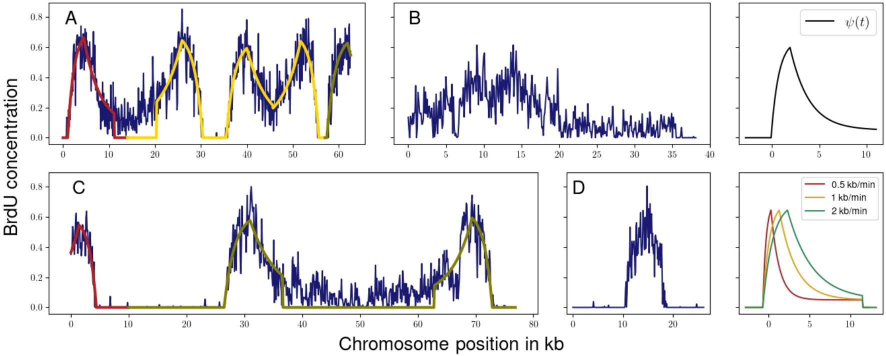


$$\min_{x} \quad \|z - Dx\|_2^2$$

$$\text{s.t} \quad \{x : \|x\|_0 \leq C\}$$

Dictionary approach

➔ The dilatation is able to detect local speed for some signals

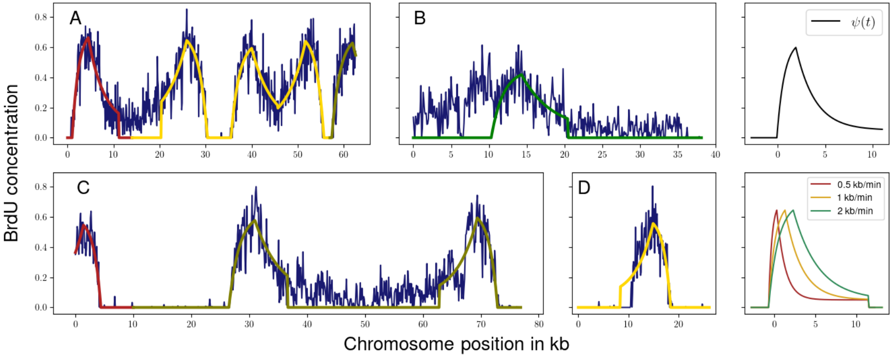


$$\min_{x} \quad \|z - Dx\|_2^2$$

$$\text{s.t} \quad \{x : \|x\|_0 \leq C\}$$

Dictionary approach

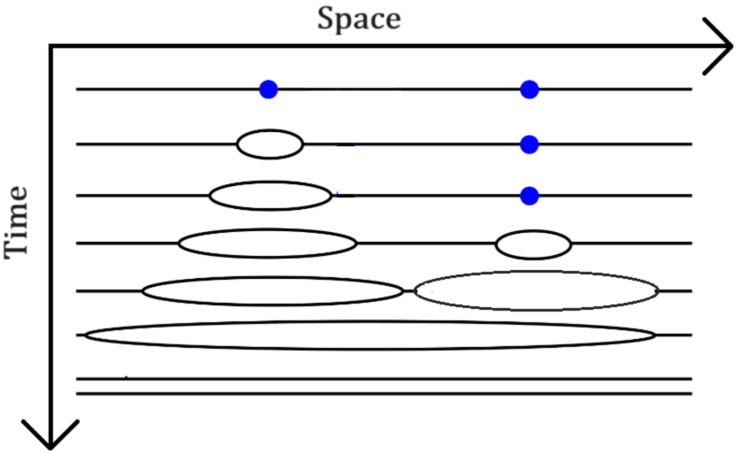
➔ The method does not cover all signals



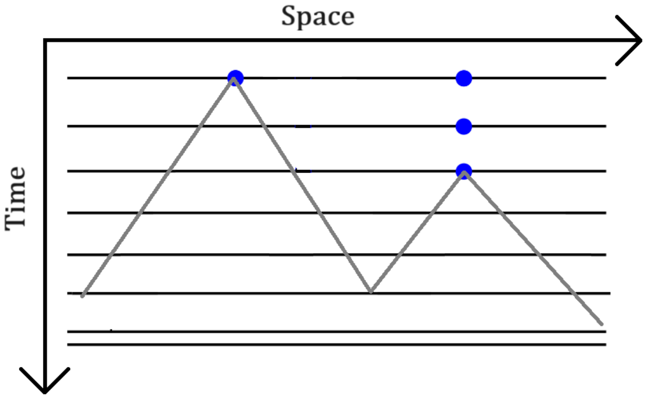
$$\min_{x} \quad \|z - Dx\|_2^2$$

$$\text{s.t} \quad \{x : \|x\|_0 \leq C\}$$

Back to the basis

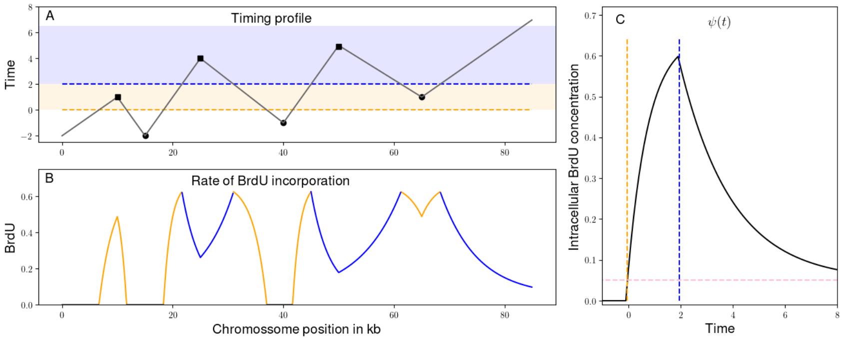


Timing profile



Assumption: Speed of replication is constant between an origin and a terminus

Change of perspective



τ - Space $\times$ Time, Piecewise linear function

ψ - Time $\times$ BrdU, concentration function

z - Space $\times$ BrdU, Signal: $z(x) = \psi(\tau(x))$

The DNA replication as a nonlinear inverse problem

- ➔ Set of equidistant points $\mathcal{X} = \{x_1, ..., x_n\}$
- ➔ Timing profile: $\boldsymbol{\tau} = (\tau_1, ..., \tau_n) := (\tau(x_1), ..., \tau(x_n))$
- ➔ Measurement operator:

$$\begin{aligned} \boldsymbol{\Psi} : \mathbb{R}^n &\longrightarrow \mathbb{R}^n_+ \\ \boldsymbol{\tau} &\mapsto (\psi(\tau_1), ..., \psi(\tau_n)) \end{aligned}$$

Consider the inverse problem:

$$\min_{\boldsymbol{\tau} \in \mathcal{P}_C} \|\mathbf{z} - \boldsymbol{\Psi}(\boldsymbol{\tau})\|_2^2, \qquad \textbf{(P1)}$$

where:

$$\mathcal{P}_C := \{\boldsymbol{\tau} \ : \ \|\mathrm{L}\boldsymbol{\tau}\|_0 \leq C\}, \quad \mathrm{L}\boldsymbol{\tau} = \ell * \boldsymbol{\tau}, \text{ with } \ell = [1, -2, 1].$$

$\mathcal{P}_C$ is the set of piecewise linear vectors with at most C breaks.

Nonlinear inverse problem

Challenges

$$\widehat{\boldsymbol{\tau}} := \arg\min_{\boldsymbol{\tau}\in\mathcal{P}_C} \|\mathbf{z} - \Psi(\boldsymbol{\tau})\|_2^2 \qquad \textbf{(P.1)}$$

- Ψ is nonlinear
- Ψ is not injective

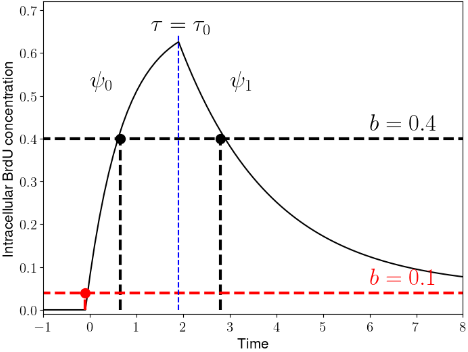
Objective: Provide a global solutions to problem (**P1**)

Assumption

- The form of the operator Ψ is important
- $\Psi(\boldsymbol{\tau}) = (\psi(\tau_1), ..., \psi(\tau_n))$

A.1 : $\exists\ \tau_0 > 0$ such that:

- $\psi_0 := \psi|_{[0,\tau_0]}$ is concave
- $\psi_1 := \psi|_{[\tau_0,\infty)}$ is convex
- The convexity or concavity of ψ_0 or ψ_1 is strict.
- both ψ_0, ψ_1 are injective.



As a consequence: $\#\psi^{-1}(b) = \#\{\boldsymbol{\tau} : \psi(\boldsymbol{\tau}) = b\} \leq 2, \quad \forall b \in \mathbb{R}_+$

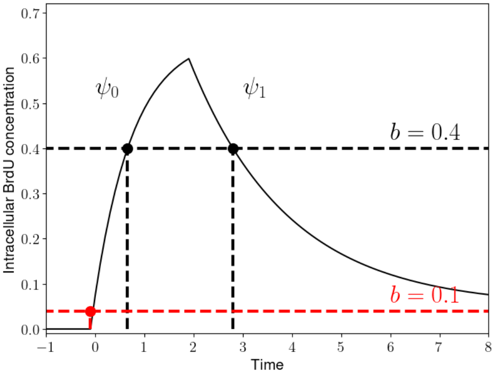
Alternative formulation

If Ψ is not injective, the definition of **P.2** is not intuitive.

$$\min_{\tau \in \mathcal{P}_C} \|\mathbf{z} - \Psi(\tau)\|_2^2 \quad \textbf{(P.1)} \quad \longleftrightarrow \quad ? \quad \textbf{(P.2)}$$

➔ Use the property: $\#\psi^{-1}(b) \leq 2, \quad \forall b \in \mathbb{R}_+$ to enumerate the inverse set

Inverse image



- Inverse image of ψ for $b \in \mathbb{R}_+$:

$$\psi^{-1}(b) = \psi_0^{-1}(b) \cup \psi_1^{-1}(b)$$

- Inverse image of Ψ for $\boldsymbol{d} \in \{0,1\}^n$:

$$\Psi_{\boldsymbol{d}}^{-1}(\mathbf{z}) := \psi_{d_0}^{-1}(\mathrm{z}_0) \times \ldots \times \psi_{d_n}^{-1}(\mathrm{z}_n)$$

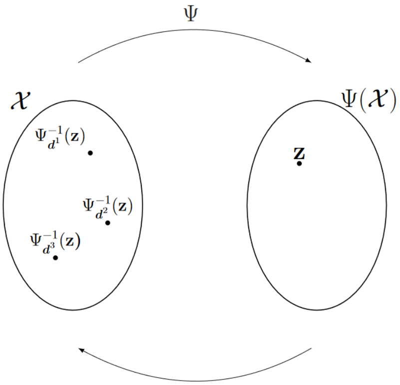
- Define:

$$\mathcal{K}(\mathbf{z}) = \left\{ \boldsymbol{d} \in \{0,1\}^n : \psi_{d_i}^{-1}(\mathrm{z}_i) \neq \varnothing \right\}$$

- Then: $\Psi_{\boldsymbol{d}}^{-1}(\mathbf{z}) \in \mathbb{R}^n$, for $\boldsymbol{d} \in \mathcal{K}(\mathbf{z})$

$$\boxed{\Psi(\boldsymbol{\tau}) = (\psi(\tau_1), \psi(\tau_2), ..., \psi(\tau_n))}$$

Taylor approximation



Because of Taylor approximation theorem, for each $\boldsymbol{d} \in \mathcal{K}(\mathbf{z})$

$$\|\mathbf{z} - \Psi(\boldsymbol{\tau})\|_2^2 \approx \sum_{i=1}^{n} w_{\boldsymbol{d},i}^2 \left(\Psi_{\boldsymbol{d}}^{-1}(\mathbf{z}) - \boldsymbol{\tau}\right)_i^2 ,$$

Taylor approximation

Because of Taylor approximation theorem, for each $\boldsymbol{d} \in \mathcal{K}(\mathrm{z})$

$$\|\mathrm{z} - \Psi(\boldsymbol{\tau})\|_2^2 \approx \sum_{i=1}^{n} w_{\boldsymbol{d},i}^2 \left(\Psi_{\boldsymbol{d}}^{-1}(\mathrm{z}) - \boldsymbol{\tau}\right)_i^2,$$

where $w_{\boldsymbol{d},i} := \psi'(\psi_{d_i}^{-1}(\mathrm{z}_i)) \in \mathbb{R}$. We define: $\|\boldsymbol{v}\|_{\boldsymbol{w}} := \sqrt{\sum_{i=1}^{n} w_i^2 v_i^2}$. Then:

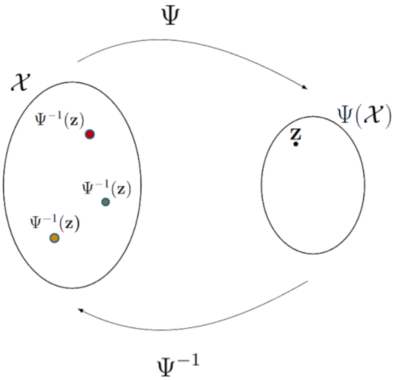
$$\text{For each } \boldsymbol{d} \in \{0,1\}^n : \quad \|\mathrm{z} - \Psi(\boldsymbol{\tau})\|_2^2 \approx \|\Psi_{\boldsymbol{d}}^{-1}(\mathrm{z}) - \boldsymbol{\tau}\|_{\boldsymbol{w}_d}^2$$

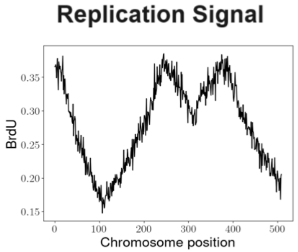
Resulting in:

$$\boxed{\min_{\{(\boldsymbol{\tau},\boldsymbol{d}) \in \mathcal{P}_C \times \{0,1\}^n\}} \|\Psi_{\boldsymbol{d}}^{-1}(\mathrm{z}) - \boldsymbol{\tau}\|_{w_{\boldsymbol{d}}}^2 \quad \textbf{(P.2)}}$$

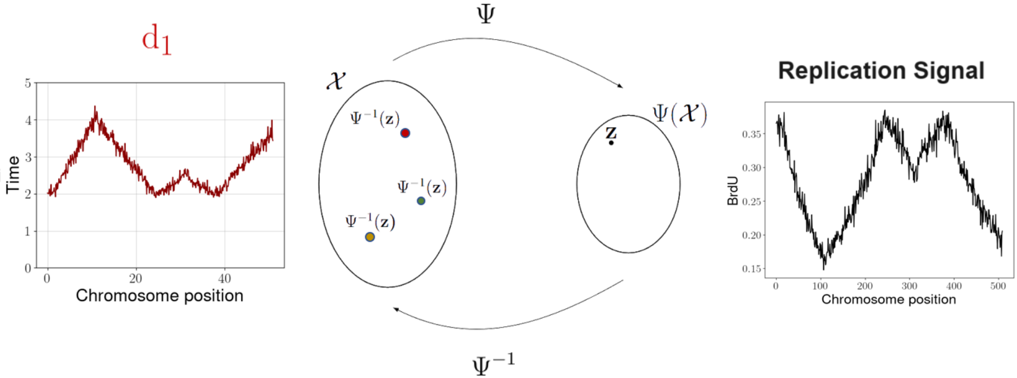
Intuition of **(P.2)** with a simulated signal

$$\Psi^{-1}(z) = \{x : \Psi(x) = z\}$$

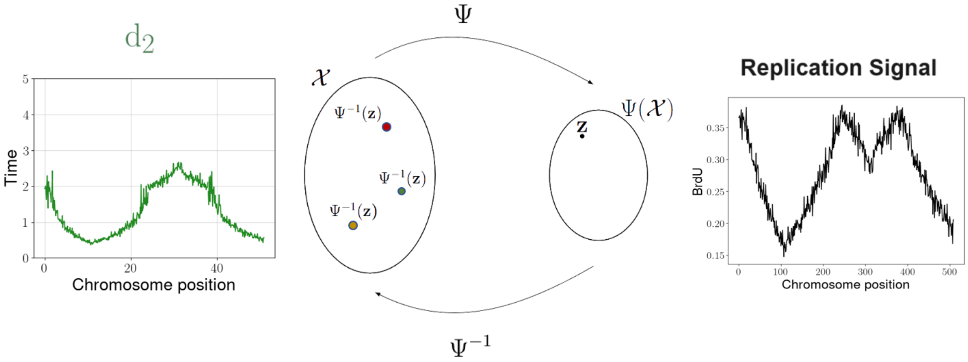




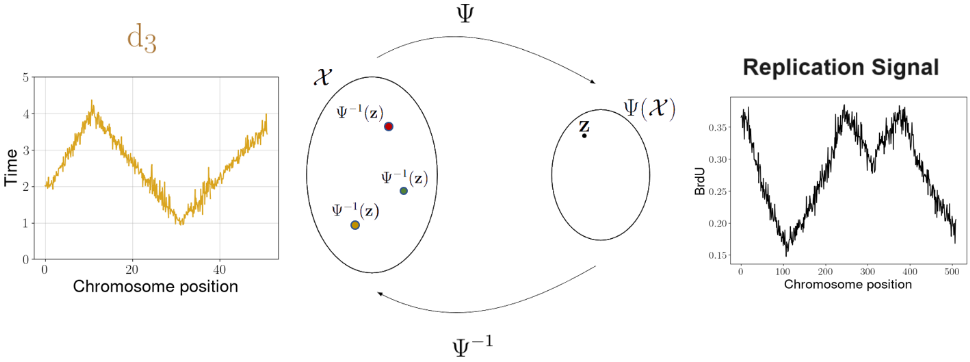
Intuition of **(P.2)** with a simulated signal



Intuition of **(P.2)** with a simulated signal



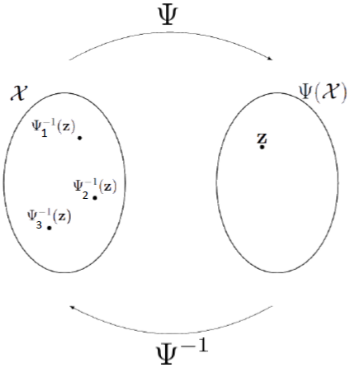
Intuition of **(P.2)** with a simulated signal

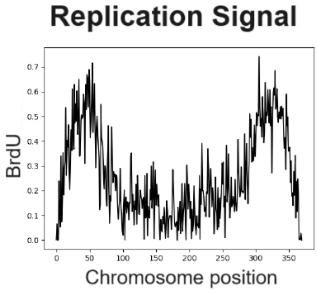


➔ One of the inverse images is close to a piecewise linear behavior

Intuition of weights w_d in a real signal

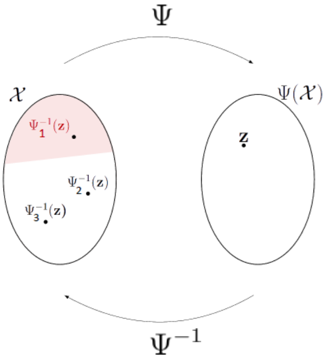
$$\Psi^{-1}(z) = \{x : \Psi(x) = z\}$$

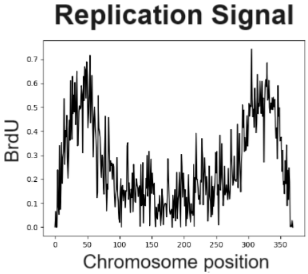




Intuition of weights w_d in a real signal

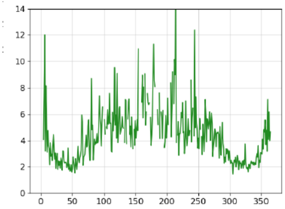
$$\Psi^{-1}(\mathrm{z}) = \{x : \Psi(x) = \mathrm{z}\}$$

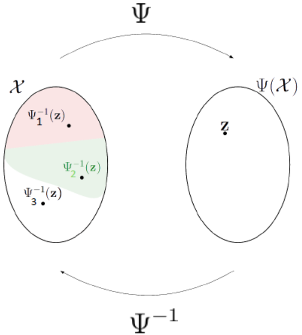


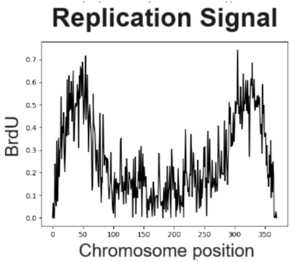


Intuition of weights w_d in a real signal

$$\Psi^{-1}(\mathrm{z}) = \{x : \Psi(x) = \mathrm{z}\}$$

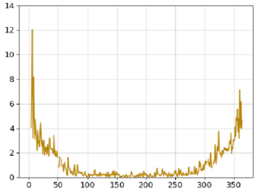


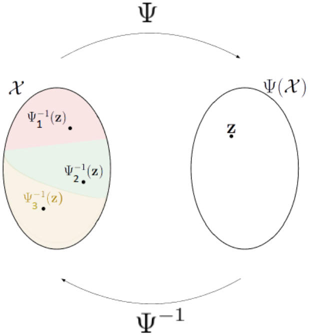


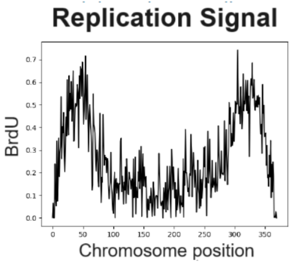


Intuition of weights $\boldsymbol{w_d}$ in a real signal

$$\Psi^{-1}(\mathrm{z}) = \{x : \Psi(x) = \mathrm{z}\}$$







➔ One of the inverse images is close to a piecewise linear behavior

Equivalence in noiseless case

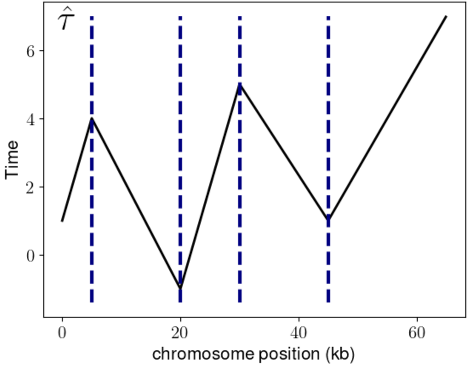
$$\hat{\boldsymbol{\tau}} := \underset{\{\boldsymbol{\tau} \in \mathcal{P}_C\}}{\arg\min} \|\mathrm{z} - \Psi(\boldsymbol{\tau})\|_2^2 \quad \textbf{(P.1)} \quad \longleftrightarrow \quad (\boldsymbol{\tau}^*, \boldsymbol{d}^*) := \underset{\{(\boldsymbol{\tau}, \boldsymbol{d}) \in \mathcal{P}_C \times \mathcal{K}(\mathrm{z})\}}{\arg\min} \|\boldsymbol{\tau} - \Psi_{\boldsymbol{d}}^{-1}(\mathrm{z})\|_{w_{\boldsymbol{d}}}^2 \quad \textbf{(P.2)}$$

Theorem. Suppose $\hat{\boldsymbol{\tau}}$ or $\boldsymbol{\tau}^*$ do not have constant parts and its breakpoints are sufficiently spaced (more then 1.2 kb). Then $\hat{\boldsymbol{\tau}} = \boldsymbol{\tau}^*$, and $\boldsymbol{d}^*$ is such that $\Psi_{\boldsymbol{d}^*}^{-1}(\mathrm{z}) = \boldsymbol{\tau}^* = \hat{\boldsymbol{\tau}}$

Proof.
Based on the injectivity of Ψ restricted to the correspondent set in $\mathcal{P}_C$. □

Equivalence in noiseless case

Theorem. Suppose $\hat{\boldsymbol{\tau}}$ or $\boldsymbol{\tau}^*$ do not have constant parts and its breakpoints are sufficiently spaced (more then 1.2 kb). Then $\hat{\boldsymbol{\tau}} = \tau^*$, and $\boldsymbol{d}^*$ is such that $\Psi_{\boldsymbol{d}^*}^{-1}(\mathrm{z}) = \boldsymbol{\tau}^* = \hat{\boldsymbol{\tau}}$



Numerical method

Rewriting **(P.2')**

$$\min_{\{(\boldsymbol{\tau},\boldsymbol{d})\in\mathcal{P}_C\times\{0,1\}^n\}} \|\Psi_{\boldsymbol{d}}^{-1}(\mathrm{z}) - \boldsymbol{\tau}\|_{w_{\boldsymbol{d}}}^2 \quad \textbf{(P.2)}$$

$=$ ($\odot$ coordinatewise multiplication)
$\boldsymbol{w_d} = \boldsymbol{d} \odot \boldsymbol{w_1} + (\boldsymbol{1} - \boldsymbol{d}) \odot \boldsymbol{w_0}$

$$\begin{aligned} \min_{\boldsymbol{\tau},\boldsymbol{d}} \quad & \tfrac{1}{2}\|\boldsymbol{d} \odot (\boldsymbol{\tau} - \Psi_{\boldsymbol{1}}^{-1}(\mathrm{z}))\|_{\boldsymbol{w_1}}^2 + \tfrac{1}{2}\|(\boldsymbol{1} - \boldsymbol{d}) \odot (\boldsymbol{\tau} - \Psi_{\boldsymbol{0}}^{-1}(\mathrm{z}))\|_{\boldsymbol{w_0}}^2 \\ \text{s.t.} \quad & \boldsymbol{\tau} \in \mathbb{R}^n, \ \|\mathrm{L}\boldsymbol{\tau}\|_0 \le C \\ & \boldsymbol{d} \in \{0,1\}^n \end{aligned}$$

$\approx$ (ℓ_1 regularization) $\lambda > 0$

$$\begin{aligned} \min_{\boldsymbol{\tau},\boldsymbol{d}} \quad & \tfrac{1}{2}\|\boldsymbol{d} \odot (\boldsymbol{\tau} - \Psi_{\boldsymbol{1}}^{-1}(\mathrm{z}))\|_{\boldsymbol{w_1}}^2 + \tfrac{1}{2}\|(\boldsymbol{1} - \boldsymbol{d}) \odot (\boldsymbol{\tau} - \Psi_{\boldsymbol{0}}^{-1}(\mathrm{z}))\|_{\boldsymbol{w_0}}^2 + \lambda\|\mathrm{L}\boldsymbol{\tau}\|_1 \quad \textbf{(P.3)} \\ \text{s.t.} \quad & \boldsymbol{\tau} \in \mathbb{R}^n, \\ & \boldsymbol{d} \in \{0,1\}^n \end{aligned}$$

Solving the regularized problem

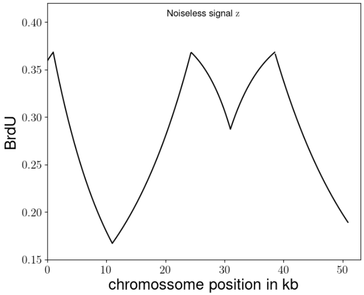
$$\begin{aligned}\min_{\boldsymbol{\tau},\boldsymbol{d}} \quad & \tfrac{1}{2}\|\boldsymbol{d} \odot (\boldsymbol{\tau} - \Psi_{\mathbf{1}}^{-1}(\mathbf{z}))\|_{\boldsymbol{w}_1}^2 + \tfrac{1}{2}\|(\mathbf{1} - \boldsymbol{d}) \odot (\boldsymbol{\tau} - \Psi_{\mathbf{0}}^{-1}(\mathbf{z}))\|_{\boldsymbol{w}_0}^2 + \lambda\|\mathrm{L}\boldsymbol{\tau}\|_1 \qquad \textbf{(P.3)} \\ \text{s.t.} \quad & \boldsymbol{\tau} \in \mathbb{R}^n, \\ & \boldsymbol{d} \in \{0,1\}^n\end{aligned}$$

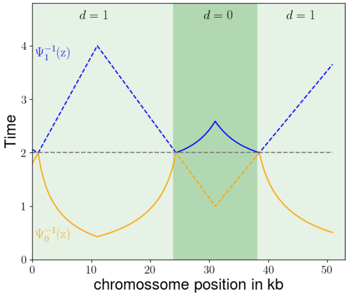
- ➔ For each $\boldsymbol{d} \in \{0,1\}^n$, **(P.3)** is convex
- ➔ For each $\boldsymbol{d} \in \{0,1\}^n$, **(P.3)** is similar to *generalized lasso*

Proposition. For each $\boldsymbol{d} \in \{0,1\}^n$, problem **(P.3)** is equivalent to its dual formulation, which is a *quadratic optimization problem.*

Constraining the set $\{0,1\}^n$

- The set $\{0,1\}^n$ is excessively large to be tractable
- We can constraint this set without changing its optimal solution





$$\mathcal{D} = \{\boldsymbol{d} : \boldsymbol{d} \text{ can be the optimal solution}\} = \{\boldsymbol{d} : d_i = d_{i+1},\ \forall i \in I_{\mathcal{A}}\}$$

$$I_{\mathcal{A}} = \{i : |\psi_0^{-1}(z_i) - \psi_1^{-1}(z_i)| > \epsilon\}$$

Algorithm

Algorithm: DNA-Inverse

Data: Input data: $\mathbf{z}$. Parameters: $\epsilon > 0$.

Initialization:

Compute weights $w_{\boldsymbol{d}}$. Compute set $\mathcal{D}$. Set: $\mathcal{D}_{\text{past}} = \varnothing$;

Main Loop:

for $\boldsymbol{d} \in \mathcal{D}$ **do**

- **Step 0:** $\mathcal{D}_{\text{past}} \leftarrow \mathcal{D}_{\text{past}} \cup \{\boldsymbol{d}\}$;
- **Step 1:** Solve **(P.3)**, obtaining a solution $\boldsymbol{\tau}_{\boldsymbol{d}}^*$;
- **Step 2:**

$$\boldsymbol{d}^* := \underset{\boldsymbol{d} \in \mathcal{D}_{\text{past}}}{\arg\min}\, F(\boldsymbol{\tau}_{\boldsymbol{d}}^*), \quad \boldsymbol{\tau}^* := \tau_{\boldsymbol{d}^*}^*,$$

Output: τ^*, d^*

---

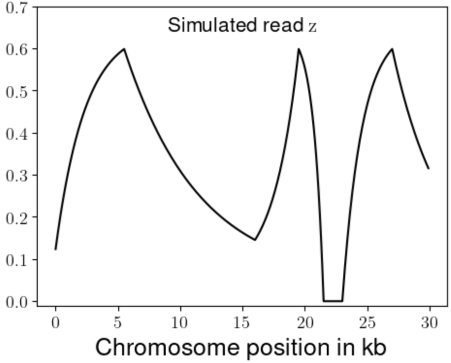
where:

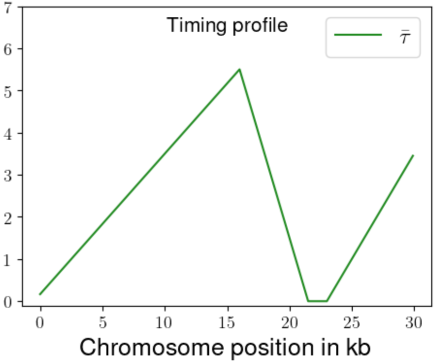
$$F(\boldsymbol{\tau}) := \frac{1}{2}\|\boldsymbol{d} \odot (\boldsymbol{\tau} - \Psi_{\mathbf{1}}^{-1}(\mathrm{z}))\|_{\boldsymbol{w}_1}^2 + \frac{1}{2}\|(\mathbf{1} - \boldsymbol{d}) \odot (\boldsymbol{\tau} - \Psi_{\mathbf{0}}^{-1}(\mathrm{z}))\|_{\boldsymbol{w}_0}^2$$

Numerical results

Methods for nonlinear inverse problem

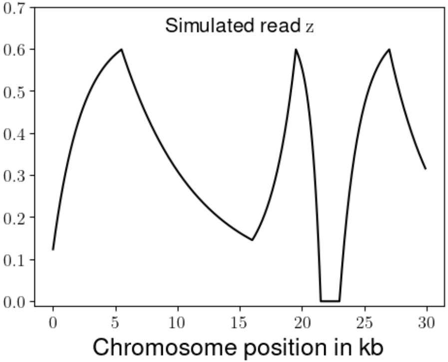
➔ Example of a noiseless signal

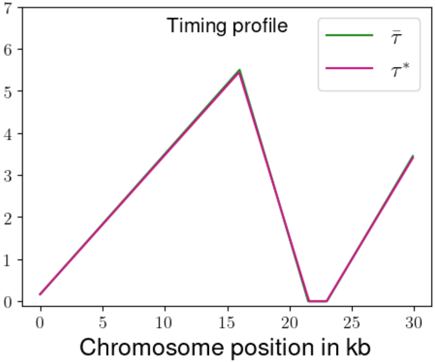




Methods for nonlinear inverse problem

➔ DNA Inverse solution (pink)





Methods for nonlinear inverse problem

- Other optimization methods can solve the ℓ_1-regularized (P.1)

$$\min_{\boldsymbol{\tau}\in\mathbb{R}^n} \|\mathbf{z} - \Psi(\boldsymbol{\tau})\|_2^2 + \gamma\|\mathrm{L}\boldsymbol{\tau}\|_1,$$

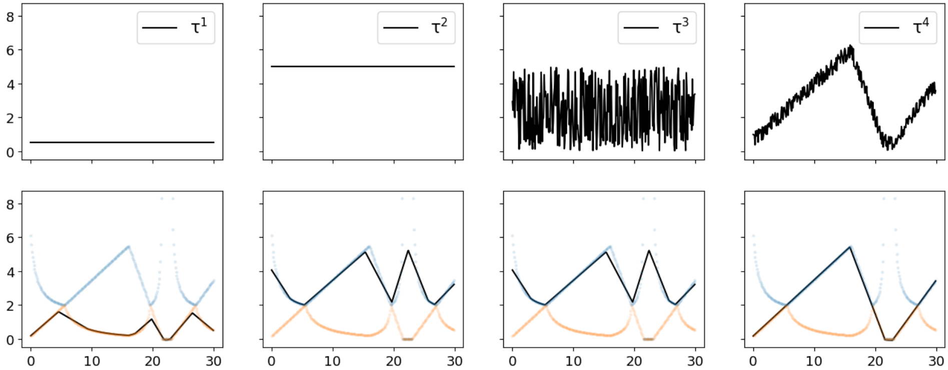
- One of these methods is the Primal-Dual method (Valkonen, 2019):

$$G(\boldsymbol{u}) = \|\boldsymbol{u} - \mathbf{z}\|_2^2, \quad G^*(\boldsymbol{y}) = \sup_{\boldsymbol{u}\in\mathbb{R}^n} \langle \boldsymbol{u}, \boldsymbol{y}\rangle - G(\boldsymbol{u}).$$

$$\begin{cases} \boldsymbol{\tau}^{k+1} = \mathrm{prox}_{\sigma_1\gamma\|\cdot\|_1}(\boldsymbol{\tau}^k - \sigma_1\Psi'(\boldsymbol{\tau}^k)y^k) \\ \boldsymbol{y}^{k+1} = \mathrm{prox}_{\sigma_2(G^* - 2\langle\Psi(\boldsymbol{\tau}^k),\cdot\rangle)}(\boldsymbol{y}^k - \sigma_2\Psi(\boldsymbol{\tau}^k)), \end{cases}$$

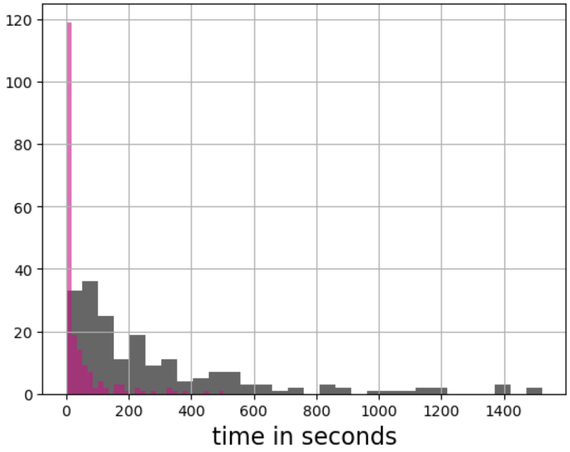
Methods for nonlinear inverse problem

➔ The PDPS provide local solutions that depend on the initial point: $\boldsymbol{\tau}^i, i \in \{1, 2, 3, 4\}$



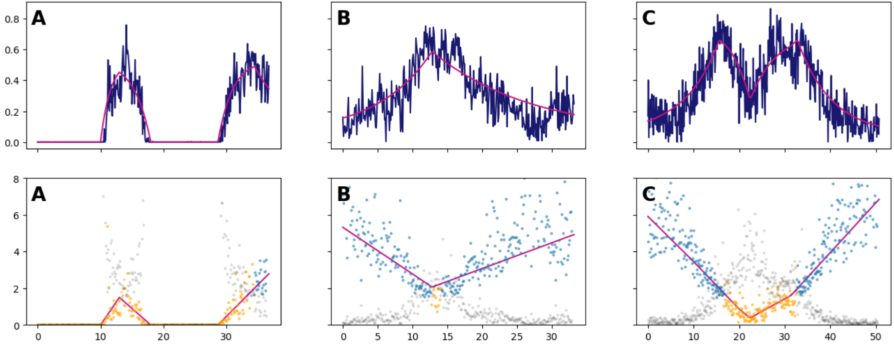
Execution time

➔ When compared to PDPS with smart initialization (grey), DNA inverse (pink) exhibits faster performance



Results for real data

➔ DNA inverse identifies the biological reality behind the signal



➔ Timing profile $\boldsymbol{\tau}^*$ (below, pink), and a signal approximation (above,pink)

➔ The colored dots are selected by the integer variable $\boldsymbol{d}^*$

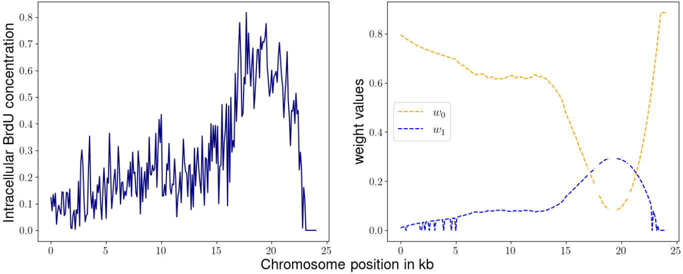
Thanks for your attention!!

References

Valkonen, T. et al (2019). “Acceleration and Global Convergence of a First-Order Primal-Dual Method for Nonconvex Problems”. In: *SIAM Journal on Optimization* 29.1, pp. 933–963. DOI: 10.1137/18M1170194. URL: https://doi.org/10.1137/18M1170194.

Extension to the noisy case

- ➔ We extend $\boldsymbol{w_d}$ to be zero in coordinates $i \in \{1, \ldots, n\}$ such that $\psi_{d_i}(\mathbf{z}) = \emptyset$
- ➔ These coordinates contain less information about the signal position



In the noisy case, consider the optimization problem:

$$(\boldsymbol{\tau}^*, \boldsymbol{d}^*) := \underset{\{(\boldsymbol{\tau},\boldsymbol{d}) \in \mathcal{P}_C \times \{0,1\}^n\}}{\arg\min} \|\Psi_{\boldsymbol{d}}^{-1}(\mathbf{z}) - \boldsymbol{\tau}\|_{w_{\boldsymbol{d}}}^2 \quad \textbf{(P.2')}$$