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

Clara Lage, Nelly Pustelnik, Jean-Michel Arbona, Benjamin Audit, Rémi Gribonval

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

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

[1] Univ Lyon, ENS de Lyon, CNRS, Laboratoire de Physique, F-69342 Lyon, France; [2] Laboratoire de Biologie et Modélisation de la Cellule, ENS de Lyon, Lyon, France; [3] Univ Lyon, ENS de Lyon, Inria, CNRS, UCBL, LIP UMR 5668, F-69342 Lyon, France.

Abstract

DNA replication stands as one of the fundamental biological processes crucial for cellular functioning. Recent experimental developments enable the study of replication dynamics at the single-molecule level for complete genomes, facilitating a deeper understanding of its main parameters. In these new data, replication dynamics is reported by the incorporation of an exogenous chemical, whose intra-cellular concentration follows a nonlinear function. The analysis of replication traces thus gives rise to a nonlinear inverse problem, presenting a nonconvex optimization challenge. We demonstrate that under noiseless conditions, the replication dynamics can be uniquely identified by the proposed model. Computing a global solution to this optimization problem is specially challenging because of its multiple local minima. We present the DNA-inverse optimization method that is capable of finding this global solution even in the presence of noise. Comparative analysis against state-of-the-art optimization methods highlights the superior computational efficiency of our approach. DNA-inverse enables the automatic recovery of all configurations of the replication dynamics, which was not possible with previous methods.

1 Introduction.

Context DNA replication is the cellular process by which a cell makes an identical copy of all its chromosomes. It is a highly parallelized DNA synthesis process under strong biological regulation. Its successful completion at each cell cycle is crucial to ensure that genetic information is accurately passed on from one generation to the next. Genetic diseases can appear from replication errors in the germline while genetic instabilities associated to perturbations of the replication dynamic is a recurrent pattern in the appearance and progression of cancer [1]. Hence, the characterization of the so-called *DNA replication program* is not only of fundamental interest but also as implication on human health.

The replication program for one cell can be described by the replication time versus chromosome position curve: $\tau(x)$ (Figure 1A). Single-molecule experimental characterization techniques (i) submit the cells to a pulse of a modified nucleotide called BrdU so that the intracellular BrdU concentration follows a time pulse $\psi(t)$ (Figure 1C) and (ii) measure a posteriori the resulting BrdU incorporation profile $z(x)$ along single DNA molecules (Figure 1B). The task of characterizing the DNA replication program thus consists in inferring $\tau(x)$ from $z(x)$ given $\psi(t)$ by solving the inverse problem $z(x) = \psi(\tau(x))$. Following previous work in the field, we assume that the rate of DNA synthesis is locally constant, i.e., that $\tau(x)$ is piecewise linear. This configurations results in a non-linear inverse problem defined over the set of function with sparse second derivative.

Nonlinear inverse problems are a field in expansion and with significant applications in imagery, optics, and tomography. [2, 3, 4, 5, 6]. In contrast to linear inverse problems, there is no optimization method capable of providing global solutions to large classes of problems [7, 8]. On the other hand, recent works have successfully adapted methods used in linear inverse problems, such as proximal methods and ADMM, to find local solutions in a nonlinear framework [9, 5]. The difficulty of

transitioning from a local solution to a global one stems from the fact that a nonlinear measurement operator, as in the case of DNA replication, often corresponds to a nonconvex optimization problem. In this work, we propose to tackle a nonlinear inverse problem by an approximation that yields a mixed-integer nonlinear programming problem (MINLP). The approach of solving nonconvex optimization problems by introducing integer variables is present in recent literature, particularly in the context of nonconvex machine learning problems [10, 11] and to replace non-convex constraints [12]. In our case study, the integer variable enables us to decompose a nonconvex problem into a family of convex problems.

State-of-the-art The *replication timing profile* $\tau(x)$ depends on the location and time of activation of the so-called *replication origins* and the speed of the *replication forks* (Figure 1A). The experimental signals obtained by FORK-seq [13, 14] captures the DNA synthesis by measuring the variation of the concentration of BrdU, a modified nucleotide that incorporates in replacement of thymidines along a fragment of chromosome, called a *read* (Figure 1). Typical examples of experimental signals are illustrated in Figure 2. The signal resulting from one fork is an atom having the shape of the BrdU time pulse $\psi(t)$ with a spatial dilatation depending on the local replication velocity (Figure 2F). In previous studies [14, 13, 15, 16], signal processing methods applied to FORK-seq data enabled to estimate the position, orientation, and speed of DNA replication forks, but only in replication fork configurations where fork atoms are sufficiently isolated (Figure 2A,D). In [14], the numerical approach involves a piecewise linear approximation of the BrdU vs. space signal. In [15, 16] the function ψ is used as a reference atom in a dictionary composed by translation and rescaling of ψ (Figure 2F) leading to a sparse coding approach. According to [15], the most effective numerical method for sparse coding in signals with high noise is Matching Pursuit [17].

While successful in accurately estimating fork speed, these approaches fail to characterize replication motifs involving truncated atoms that appears in the vicinity of replication origins and termini (Figure 1B). Indeed, since fork progression start at origins and ends when converging forks merge (each loci is replicated once and only once), the significant particularity of the BrdU incorporation signals is that the contributions of distinct forks never overlap and add up, fork atoms being truncated at replication origins and termini (Figure 1B). Therefore, there is a need for an alternative approach capable of robustly extracting these diverging or converging fork configurations (see Figure 2B,E). In this context, we move away from the additive atom approach and instead focus on a method that can determine the time profile directly by specifying a nonlinear inverse problem.

Contribution and outline The contribution of this article is twofold. We introduce, for the first time, a nonlinear inverse problem that accurately models the biological configuration of FORK-seq data. We thoroughly investigate the theoretical aspects of the model and demonstrate that the optimization problem yields a unique solution under specific assumptions. On the other side, we propose an original numerical method capable of globally solving this problem, a task that is recent investigated in the existing literature [18, 4, 6].

The article is divided as follows: In Section 2 we introduce the DNA-inverse model that is able to capture the main aspects of the biological configuration that results in FORK-seq data as well as the nonlinear inverse problem that provides the timing profile associated to a signal. The theoretical analysis of this problem is provided in Section 3, where we study conditions for a unique solution. In Section 4, we reformulate the optimization problem to propose a numerical method capable of finding a global solution. The development of this numerical method and algorithm is presented in Section 5. Finally, Section 6 presents the numerical results and compares it with results obtained using a state-of-the-art method.

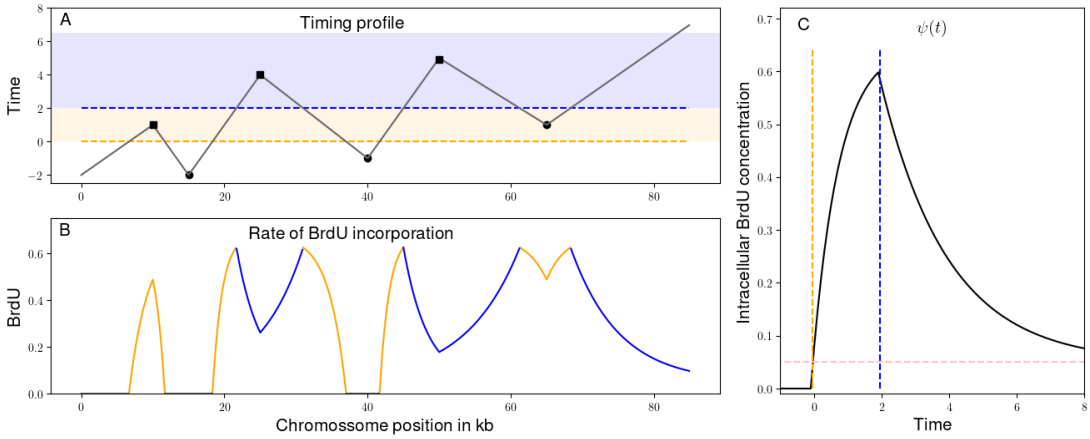

Figure 1: The DNA replication program and its characterization by pulse labeling. (A) Replication timing profile $\tau(x)$ capturing the time of replication during one replication cycle of each loci along a chromosome. DNA replication initiates at multiple sites (black dots), called *replication origins*; from each site two diverging *replication forks* emerge ensuring sequential DNA synthesis; replication terminates by the mergers of converging forks originating from neighboring origins at *replication termini* (black squares). In this example, each replication fork have a constant speed so that the timing profile between origins and termini is linear. (B) Profile $z(x)$ of the rate of BrdU incorporation along the newly synthesized DNA molecule during the replication program presented in (A) in the presence of the time pulse of BrdU $\psi(t)$ shown in (C); in the absence of noise, it is a simple composition ψ and τ: $z(x) = \psi(\tau(x))$. (C) Intracellular BrdU concentration along time, $\psi : \mathbb{R} \to \mathbb{R}_+$; the orange dashed line indicates the beginning of the external BrdU pulse resulting in the progressive increase of the intracellular BrdU concentration, the *pulse phase*; the blue dashed line indicates the end time of the external pulse (dilution of the external media) resulting in the decrease of the internal BrdU concentration, the *chase phase*. The pink dashed line represents the limit of ψ as t approaches infinity. The background color in (A) and the curve color in (B) correspond to the pulse (orange) and chase (blue) phase, respectively.

2 DNA-Inverse Model

Model: Let $\mathcal{X}$ be the discrete set of positions in a certain DNA fragment of size n. Then $\mathcal{X} := \{x_1, ..., x_n\} \subset \mathbb{R}$, where $x_{i+1} = x_i + \Delta x$, for $i \in \{1, ..., n-1\}$, and $\Delta x > 0$ is fixed as a certain distance in the chromosome scale. Assuming a locally constant speed, the motion of the molecular motor can be characterized by a piecewise linear function $\tau : \mathbb{R} \to \mathbb{R}$, that assigns to each position of the DNA fragment, the time, starting from the beginning of the experiment, at which this position have been replicated. Negative values of $\tau(x_i)$ mean that the correspondent region of the DNA fragment have been replicated before the beginning of the experiment. When restricted to positions in $\mathcal{X}$, the function $\boldsymbol{\tau}$ is associated with the vector $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)$, where $\tau_i := \tau(x_i)$, for $i \in \{1, ..., n\}$. The vector $\boldsymbol{\tau}$ is referred as a *timing profile* and is illustrated in Figure 1 A.

The measurement Ψ is defined as the coordinatewise composition of a nonlinear function ψ, which measures the concentration of BrdU in time, and the *timing profile*:

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

Denoting $\mathbf{z}$ a signal provided by FORK-seq, we suppose that there exists a timing profile $\bar{\boldsymbol{\tau}} \in \mathbb{R}^n$

such that:

$$\mathbf{z} \approx \Psi(\bar{\boldsymbol{\tau}}),$$

and that $\bar{\boldsymbol{\tau}}$ is in the set of piecewise linear vectors with a maximum of C breakpoints. This set can be expressed using the ℓ_0 pseudo-norm and a linear operator L that represents a discrete second derivative: $\mathrm{L}\boldsymbol{\tau} = \ell * \boldsymbol{\tau}$, with $\ell = [1, -2, 1]$, where $\mathrm{L} : \mathbb{R}^n \to \mathbb{R}^{n-2}$ does not consider derivatives from the borders. We denote:

$$\mathcal{P}_C := \{\boldsymbol{\tau} \ : \ \|\mathrm{L}\boldsymbol{\tau}\|_0 \leqslant C\} \tag{1}$$

for some fixed $C \in \mathbb{R}_+$. The non-negativity constraint reflects the fact that the operator Ψ returns zero for negative components of $\boldsymbol{\tau}$. In these conditions, a natural way to estimate $\bar{\boldsymbol{\tau}}$ is to solve the following optimization problem:

$$\widehat{\boldsymbol{\tau}} := \underset{\boldsymbol{\tau} \in \mathcal{P}_C}{\arg\min} \|\mathbf{z} - \Psi(\boldsymbol{\tau})\|_2^2, \tag{\textbf{P1}}$$

Nonlinear sparse coding problem: Problem (**P1**) is a *nonlinear sparse coding problem*. Nonlinear sparse coding problems appear in different application contexts such as partial differential equations, quantization and problems with large application field such as phase-retrieval [19, 20, 21, 3, 18]. When Ψ is a linear transform, problem (**P1**) can be relaxed and solved approximately by ℓ_1 regularization. The resulting optimization problem is known as the *generalized lasso* [22, 23]. In the general case, the ℓ_1 regularized version of (**P1**) fits the primal dual formulation for non-convex optimization and can be solved by a primal-dual proximal method or a generalized Alternating Direction Method of Multipliers (ADMM) for nonlinear operators Ψ [19, 24]. However, the solution proposed by these methods, as a solution to a non-convex problem, is a local solution. Additionally, these algorithms may have a high runtime depending on the difficulty in calculating the necessary proximal operators for the iterations.

Particularities of DNA-Inverse: In our case study, we present some specificities of the nonlinearity of operator Ψ. We note that the structure of the function Ψ is given coordinate-wise by the function ψ, that represents the BrdU concentration. Despite not being invertible, ψ has at most two possible antecedents ("inverses") for each point $b \in \psi(\mathbb{R}_+)$. In addition, any element of $\mathbb{R}_-$ is mapped to 0.

A.1 : $\#\psi^{-1}(b) = \#\{\boldsymbol{\tau} : \psi(\boldsymbol{\tau}) = b\} \leqslant 2, \quad \forall b \in \mathbb{R}_+$ and $\psi|_{[-\infty,0]} = 0$.

The hypothesis **A.1** is employed to develop the DNA-Inverse model (Section 4). However, to ensure uniqueness in the detected timing profile, an additional hypothesis regarding the function ψ : is necessary:

A.2 : $\exists\ \tau_0 > 0$ such that $\psi_0 := \psi|_{[0,\tau_0]}$ is convex and $\psi_1 := \psi|_{[\tau_0,\infty)}$ is concave (or the opposite), and the convexity or concavity of ψ_0 or ψ_1 is strict. In addition, both ψ_0, ψ_1 are injective.

These properties will guide the study of problem (**P1**). The existence of at most two possible inverses for ψ implies that there will be at most 2^n inverses for Ψ. In this work, we aim to narrow down the possibilities of inverses by investigating other important characteristics of the problem.

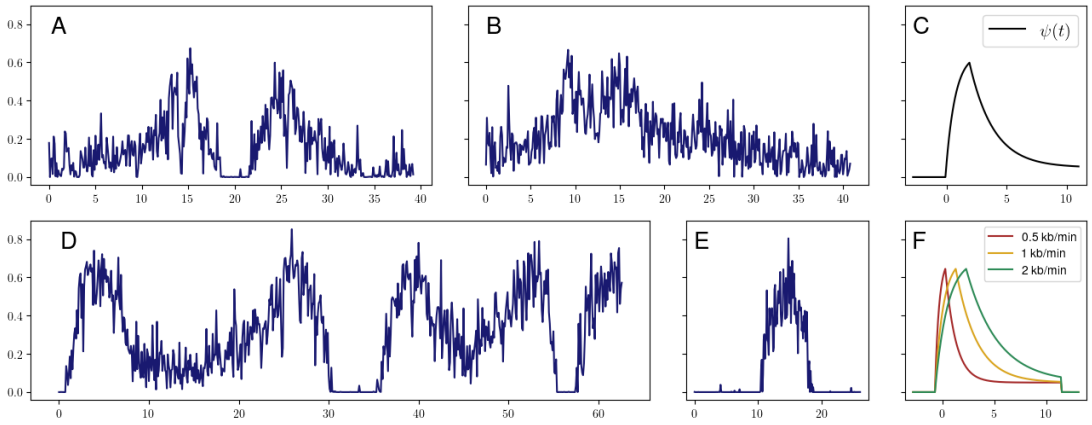

Figure 2: (A,B,D,E): examples of experimental yeast FORK-seq signals for different replication configurations. C: function $\psi(t)$ which has been experimentally determined in yeast cells [14, Section Methods]. The BrdU is injected in the medium at time $t = 0$, leading to a progressive increase of the intra-cellular concentration of BrdU, called *pulse phase*. After a short period (2 min) the chemical is removed from the medium by dilution resulting in a decrease of the intra-cellular concentration until a certain residual BrdU level, called *chase phase*. For signals in A and D the replication starts after the start of the BrdU injection, resulting in a pattern that reproduces the function ψ with a spatial dilatation depending on the local fork speed (F). Signals in B and E capture an initiation than happened before the injection start (truncated diverging fork atoms) and a termination during the pulse phase (truncated converging fork atoms), respectively. These two configurations can not be recognized by the dictionary model [16]

3 Injectivity of Ψ

In this section, we establish the uniqueness of the solution to problem (**P1**) by introducing additional constraints in the available set. In other words, we prove the injectivity of Ψ in the set of piecewise linear signals with extra constraints specified in this section. These additional constraints are divided in two types: one for analyzing the simple case of linear signals, as detailed in Section 3.1, and the second one in the general case of piecewise linear signals, addressed in Section 3.2. The constraints will be further justified in the context of the DNA replication problem, Section 3.3. This result allows us to demonstrate that the proposed model can uniquely define a *timing profile* $\widehat{\boldsymbol{\tau}}$ that better fits the DNA replication signal $\mathbf{z}$.

We begin by presenting a Lemma that facilitates the manipulation of the set $\mathcal{P}_C$ within a continuous framework.

Lemma 1 (Continuous form of the set $\mathcal{P}_C$)**.** *Let $\boldsymbol{\tau} \in \mathcal{P}_C$ and $\mathcal{X} := \{x_1, ..., x_n\} \subset \mathbb{R}$, be a set such that $x_{i+1} - x_i = \Delta x > 0$, for all $i \in \{1, ..., n-1\}$. Define $f_\tau : [x_1, x_n] \to \mathbb{R}_+$, by $f_\tau(x_i) := \tau_i$, for all $i \in \{1, ..., n\}$, and let f_τ be linear in $I_i := [x_i, x_{i+1}]$, for $i \in \{1, ..., n-1\}$. Then, f_τ is a continuous piecewise linear function on $[x_1, x_n]$ with $p \leqslant C$ breakpoints in (x_1, x_n). In addition, the breakpoints of f_τ form a subset $\{x_{i_1}, ..., x_{i_p}\} \subset \mathcal{X}$, and the indices $\{i_1, ..., i_p\} \subset \{1, ..., n-1\}$ do not depend on $\mathcal{X}$.*

Proof. Clearly f_τ is continuous and piecewise linear. We will verify that if $\boldsymbol{\tau} \in \mathcal{P}_C$, f_τ has $p = \|\mathrm{L}\boldsymbol{\tau}\|_0$ breakpoints. First, we note that there is no breakpoint in the interior of I_i, for all $i \in \{1, ..., n-1\}$ because f_τ is linear by definition. Then, all breakpoints are contained in the set $\{x_2, ..., x_{n-1}\}$. For

$i \in \{2, ..., n-1\}$, denote $p_i := (x_i, \tau_i) \in \mathbb{R}^2$. The proof is based on the following observation:

$$(\mathrm{L}\boldsymbol{\tau})_i = 0 \Leftrightarrow p_{i-1}, p_i, p_{i+1} \text{ are colinear } \Leftrightarrow x_i \text{ is not a breakpoint of } f_\tau.$$

To see the implication ($\Rightarrow$), note that $(\mathrm{L}\boldsymbol{\tau})_i = 0$ means, by definition: $\tau_{i-1} - 2\tau_i + \tau_{i+1} = 0$, which implies

$$\tau_{i-1} - \tau_i = \tau_i - \tau_{i+1}. \tag{2}$$

Then define $m := \frac{\tau_i - \tau_{i-1}}{\Delta x}$, and $c := \tau_{i-1} - m x_{i-1}$. It is easy to see that p_{i-1}, p_i, p_{i+1} are in the graph of $y(x) = mx + c$, and then are colinear. We conclude that x_i can not be a breakpoint. To see ($\Leftarrow$), let $i \in \{2, ..., n-1\}$ be such that p_{i-1}, p_i, p_{i+1} are colinear. Therefore, there exists $y(x) = mx + c$ such that p_{i-1}, p_i, p_{i+1} are in the graph of y. In this case, it is then easy to verify that (2) holds, which means that $(\mathrm{L}\boldsymbol{\tau})_i = 0$.

Thus, the set of breakpoints of f_τ writes $\{i_1, ..., i_p\} \subset \{1, ..., n\}$, and $(\mathrm{L}\boldsymbol{\tau})_{i_j} \neq 0$, for $j \in \{1, ..., p\}$. This set does not depend on the choice of $\mathcal{X}$. Clearly, $p = \|\mathrm{L}\boldsymbol{\tau}\|_0 \leqslant C$. □

Equipped with Lemma 1, we can transition between the vector $\boldsymbol{\tau}$ and its continuous counterpart $f_{\boldsymbol{\tau}}$. This lemma will be important in the proof of the main results in this section.

To investigate the injectivity of Ψ, we consider Assumption **(A.2)** and two observations: First, since $\psi_{[-\infty,0)} = 0$, timing profiles cannot be differentiated for negative values τ_i, i.e., before the beginning of the experiment. To obtain injectivity we must consider non-negative *timing profiles* $\boldsymbol{\tau}$. We also observe that ψ is not injective. Particularly, there exists $t \in [0, \tau_0)$ and $t' \in (\tau_0, \infty)$, such that $u := \psi(t) = \psi_0(t) = \psi_1(t') = \psi(t')$. Consequently, the constant vector $\mathbf{z} = \boldsymbol{u} \in \mathbb{R}^n$ has two different optimal constant solutions: $\tau = \boldsymbol{t} \in \mathbb{R}^n$, and $\tau = \boldsymbol{t}' \in \mathbb{R}^n$. For this reason, to obtain a unique solution, we need to restrict the solution set to non-constant vectors:

$$\mathcal{P}_C^{\neq} = \{\boldsymbol{\tau} \in \mathcal{P}_C : \boldsymbol{\tau} \geqslant 0, \tau_i \neq \tau_{i+1} \text{ for all } i \in \{1, ..., n-1\}\} \subset \mathbb{R}^n_+.$$

In the next section, we show that if $\boldsymbol{\tau} \neq \boldsymbol{\tau}'$ are non-constant linear vectors in $\mathcal{P}_0^{\neq}$, then $\Psi(\boldsymbol{\tau}) \neq \Psi(\boldsymbol{\tau}')$.

3.1 Injectivity of Ψ when $\boldsymbol{\tau}$ is linear

We begin by analyzing the case $C = 0$, where the feasible set of (**P1**) is the set of non-negative linear vectors. Later, we will extend this argument to non constant piecewise linear vectors.

Lemma 2 (Injectivity of Ψ for $C = 0$)**.** *Assume **(A.2)**. Then, if $n \geqslant 6$, the function $\Psi : \mathcal{P}_0^{\neq} \to \mathbb{R}^n$ is injective.*

Proof. Consider $\boldsymbol{\tau}, \boldsymbol{\tau}' \in \mathcal{P}_0^{\neq}$ such that $\Psi(\boldsymbol{\tau}) = \Psi(\boldsymbol{\tau}')$. We will prove that $\boldsymbol{\tau} = \boldsymbol{\tau}'$.

Since $\boldsymbol{\tau}, \boldsymbol{\tau}'$ are linear and non-constant, by lemma 1, there exists $m, m', c, c' \in \mathbb{R}$, with $m, m' \neq 0$, such that $\tau_i = mx_i + c$, $\tau_i' = m'x_i + c'$ for each $x_i \in \mathcal{X}$, for $i \in \mathcal{I} := \{1, ..., n\}$. We begin by partitioning the set $\mathcal{I}$ into three disjoint parts:

$$\mathcal{I}_+ = \left\{i \in \mathcal{I} : \tau_i, \tau_i' \geqslant \tau_0\right\}, \quad \mathcal{I}_- = \left\{i \in \mathcal{I} : \boldsymbol{\tau}, \boldsymbol{\tau}' \leqslant \tau_0\right\}, \quad \mathcal{I}_{+-} = \mathcal{I} \backslash (\mathcal{I}_+ \cup \mathcal{I}_-).$$

Observe that for $i \in \mathcal{I}_+$ we have $\Psi(\boldsymbol{\tau})_i = \psi_1(\tau_i)$ and similarly with $\boldsymbol{\tau}'$. Since $\Psi(\tau) = \Psi(\tau')$ and ψ_1 is injective because of **(A.2)**, we deduce that

$$\tau_i = \tau_i', \quad \forall i \in \mathcal{I}_+.$$

The same result holds for $i \in \mathcal{I}_-$, by the injectivity of ψ_0. Since both $\boldsymbol{\tau}$ and $\boldsymbol{\tau}'$ are linear vectors, they are equal as soon as they coincide at two points, hence if $\#(\mathcal{I}_+ \cup \mathcal{I}_-) \geqslant 2$ then $\tau = \tau'$. To

complete the proof, it is thus enough to verify that $\#\mathcal{I}_{+-} \leqslant 4$. To establish this fact, we proceed by contradiction, and observe first that for each $i \in \mathcal{I}_{+-}$ we have

$$\tau_i < \tau_0 < \tau_i', \text{ or } \tau_i' < \tau_0 < \tau_i.$$

Suppose that $\mathcal{I}_{+-}$ has 5 elements. By the pigeonhole principle, at least one of the two above inequalities must be satisfied by at least 3 elements $i \in \mathcal{I}_{+-}$ as illustrated on Figure 3. Without loss of generality (up to interchanging the role of $\boldsymbol{\tau}$ and $\boldsymbol{\tau}'$) we can assume that this inequality writes $\tau_i < \tau_0 < \tau_i'$, and holds for $i \in \{i_1, i_2, i_3\} \subset \mathcal{I}_{+-}$ such that $i_1 < i_2 < i_3$, implying $x_{i_1} < x_{i_2} < x_{i_3}$. Consider $t \in (0,1)$ be such that: $x_{i_2} = tx_{i_1} + (1-t)x_{i_3}$ Consider the notation of Assumption **(A.2)**, where ψ_0 is strictly concave and ψ_1 is convex (the other possibilities where ψ_0 and ψ_1 exibit different combinations of convexity and concavity can be analyzed similarly). Since ψ_0 is strictly concave and ψ_1 is convex, the functions and $m, m' \neq 0$: $\varphi_0, \varphi_1 : [x_{i_1}, x_{i_3}] \to \mathbb{R}$ defined as:

$$\varphi_0(x) = \psi_0(mx + c), \text{ and } \varphi_1(x) = \psi_1(m'x + c'),$$

are strictly concave and convex respectively. Since $\Psi(\boldsymbol{\tau}) = \Psi(\boldsymbol{\tau}')$ the functions φ_0 and φ_1 coincide in three points x_{i_1}, x_{i_2} and x_{i_3}. As φ_0 is strictly concave, we get:

$$\varphi_1(x_{i_2}) = \varphi_0(x_{i_2}) < t\varphi_0(x_{i_1}) + (1-t)\varphi_0(x_{i_3}) = t\varphi_1(x_{i_1}) + (1-t)\varphi_1(x_{i_3}),$$

which contradicts the convexity of φ_1. We conclude that $\#\mathcal{I}_{+-} \leqslant 4$ consequently $\#(\mathcal{I}_+ \cup \mathcal{I}_-) \geqslant 2$ and $\tau = \tau'$. □

Corollary 1 (Extension to the continuous case)**.** *Let $I \subset \mathbb{R}$ be an interval. Consider two linear and non-constant functions $f_1, f_2 : I \to \mathbb{R}_+$, and the compositions $\varphi^{(1)}, \varphi^{(2)} : I \to \mathbb{R}$, $\varphi^{(1)}(x) = \psi(f_1(x))$ and $\varphi^{(2)}(x) = \psi(f_2(x))$. Let $\mathcal{X} = \{x_1, x_2, ..., x_n\}$ be a set of equidistant points, and $n \geqslant 6$. Suppose that $\varphi^{(1)}(x_i) = \varphi^{(2)}(x_i)$, for $i \in \{1, ..., n\}$, then $f_1 = f_2$.*

Proof. Consider $\boldsymbol{\tau}^1, \boldsymbol{\tau}^2 \in \mathbb{R}^n$ defined by: $\tau_i^1 := f_1(x_i)$ and $\tau_i^2 = f_2(x_i)$. Then, by assumption, $\boldsymbol{\tau}^1 = \boldsymbol{\tau}^2$ and $\Psi(\boldsymbol{\tau}^1)_i = \psi(\tau_i^1) = \psi(\tau_i^2) = \Psi(\boldsymbol{\tau}^2)_i$, for all $i \in \{1, ..., n\}$, implying $\Psi(\boldsymbol{\tau}^1) = \Psi(\boldsymbol{\tau}^2)$. Since f_1, f_2 are linear non-constant, $\boldsymbol{\tau}^1, \boldsymbol{\tau}^2$ do not have constant parts, implying that $\boldsymbol{\tau}^1, \boldsymbol{\tau}^2 \in \mathcal{P}_0^{\neq}$. By Lemma 2, $\boldsymbol{\tau}^1 = \boldsymbol{\tau}^2$, and since f_1, f_2 coincide in more than two points, they define the same line and $f_1 = f_2$. □

3.2 The injectivity of Ψ when $\boldsymbol{\tau}$ is piecewise linear

In the case where $C > 0$, the feasible set of problem (**P1**) consists of piecewise linear vectors with less than C breaks. In this case, in addition to the constraint that prevents constant vectors in $\mathcal{P}_0^{\neq}$, we need to investigate the distance between two consecutive breaks of $\boldsymbol{\tau}$. The intuition of this investigation is that if arbitrarily close breaks are allowed, it is possible to oscillate around τ_0 obtaining the same image.

Definition 1 (Vector of breaks $\boldsymbol{i}^{\tau}$)**.** *Let $\boldsymbol{\tau} \in \mathbb{R}^n$, consider the indexes $\{i_1, ..., i_p\}$ for $p \leqslant C$ defined in Lemma 1. Then we define the vector of breaks of $\boldsymbol{\tau}$ by: $\boldsymbol{i}^{\boldsymbol{\tau}} := (i_0, i_1, ..., i_p, i_{p+1})$, where $i_0 = 1$ and $i_{p+1} = n$.*

Proposition 1 (Injectivity of Ψ for $C > 0$)**.** *Let ψ be as in Assumption (**A.2**). Consider:*

$$\mathcal{P}_C^{\geqslant} = \left\{ \boldsymbol{\tau} \in \mathcal{P}_C^{\neq} : \boldsymbol{i}_{k+1}^{\boldsymbol{\tau}} - \boldsymbol{i}_k^{\boldsymbol{\tau}} \geqslant 12, \text{ for all } k \in \{1, ..., p+1\}, p = |\boldsymbol{i}^{\boldsymbol{\tau}}| \right\}, \tag{3}$$

where $\boldsymbol{i}^{\boldsymbol{\tau}}$ is defined in Definition 1. Then $\Psi : \mathcal{P}_C^{\geqslant} \to \mathbb{R}^n$ is injective.

Proof. Consider $\boldsymbol{\tau}, \boldsymbol{\tau}' \in \mathcal{P}_C^{\geqslant}$, and a set of equidistant points $\mathcal{X} = \{x_1, ..., x_n\}$. Then by Lemma 1 there exists piecewise linear functions $f_{\boldsymbol{\tau}}, f_{\boldsymbol{\tau}'} : [x_1, x_n] \to \mathbb{R}_+$, where $f_{\boldsymbol{\tau}}(x_i) = \tau_i$, and $f_{\boldsymbol{\tau}'}(x_i) = \tau_i'$, for all $i \in \{1, ..., n\}$. To demonstrate that $\boldsymbol{\tau} = \boldsymbol{\tau}'$, it is sufficient to show that $f_{\boldsymbol{\tau}} = f_{\boldsymbol{\tau}'}$

By Lemma 1, breakpoints of $f_{\boldsymbol{\tau}}$ and $f_{\boldsymbol{\tau}'}$ are given by $\mathcal{X}^{\boldsymbol{\tau}} := \{x_{i_2}, ..., x_{i_p}\}$ and $\mathcal{X}^{\boldsymbol{\tau}'} := \{x_{i_2'}, ..., x_{i_{p'}'}\}$ for $p, p' \leqslant C$. Consider the vector $\boldsymbol{b}$, which aggregates and sorts all breakpoints in $\mathcal{X}^{\boldsymbol{\tau}} \cup \mathcal{X}^{\boldsymbol{\tau}'}$. We divide the interval $[x_1, x_n]$ into intervals $\mathcal{B}_l = [b_l, b_{l+1}]$, for $l \in \{1, ..., L-1\}$. Note that $f_{\boldsymbol{\tau}}|_{\mathcal{B}_l}$ and $f_{\boldsymbol{\tau}'}|_{\mathcal{B}_l}$ are both linear by the definition of $\boldsymbol{b}$. Consider $\mathcal{I}_l := \mathcal{B}_l \cap \mathcal{X}$. Then, by Corollary 1, for all $l \in \{1, ..., L\}$:

$$\#\mathcal{I}_l \geqslant 6 \Rightarrow f_{\boldsymbol{\tau}}|_{\mathcal{B}_l} = f_{\boldsymbol{\tau}'}|_{\mathcal{B}_l}, \tag{4}$$

Note that, according to the structure of $\mathcal{P}_C^{\geqslant}$, since $\boldsymbol{i}^{\boldsymbol{\tau}}, \boldsymbol{i}^{\boldsymbol{\tau}'}$ contains the borders $\{1, n\}$, the first and last breakpoints of $f_{\boldsymbol{\tau}}$ and $f_{\boldsymbol{\tau}'}$ can not appear before 12 points, implying that $\#\mathcal{I}_1, \#\mathcal{I}_L > 6$. We proceed by investigating the case $\#\mathcal{I}_l < 6$. Define:

$$\mathcal{L} = \{l \in \{2, ..., L-1\} : \#\mathcal{I}_l < 6\}.$$

The objective is to show that for all $l \in \mathcal{L}$, $f_{\boldsymbol{\tau}}|_{\mathcal{B}_l} = f_{\boldsymbol{\tau}'}|_{\mathcal{B}_l}$. We begin by proving that $l \in \mathcal{L}$ implies $l-1 \notin \mathcal{L}$ and $l+1 \notin \mathcal{L}$ and we proceed by contradiction. By definition of $\mathcal{I}_l$, a breakpoint for either $f_{\boldsymbol{\tau}}$ or $f_{\boldsymbol{\tau}'}$ occurs in l and in $l+1$. If $l+1 \in \mathcal{L}$, an additional breakpoint emerges in $l+2$. Consequently, $f_{\boldsymbol{\tau}}$ or $f_{\boldsymbol{\tau}'}$ would accumulate two breakpoints in an interval of length less then 12 which means that τ or τ' is not in the set $\mathcal{P}_C^{\geqslant}$. The same argument applies for the adjacent interval $\mathcal{I}_{l-1}$. We conclude that $l-1$ and $l+1$ are not in $\mathcal{L}$. To complete the proof we demonstrate that:

$$\text{If } l-1, l+1 \notin \mathcal{L}, \text{ then } f_{\boldsymbol{\tau}}|_{\mathcal{B}_l} = f_{\boldsymbol{\tau}'}|_{\mathcal{B}_l}.$$

Note that adjacent intervals share boundaries, thus for every $l \in \{1, ..., L-1\}$, $\boldsymbol{b}_{l+1} \in \mathcal{B}_l \cap \mathcal{B}_{l+1}$ and $\boldsymbol{b}_l \in \mathcal{B}_l \cap \mathcal{B}_{l-1}$. If $l-1, l+1 \notin \mathcal{L}$, applying (4), we have that: $f_{\boldsymbol{\tau}}|_{\{\boldsymbol{b}_l, \boldsymbol{b}_{l+1}\}} = f_{\boldsymbol{\tau}'}|_{\{\boldsymbol{b}_l, \boldsymbol{b}_{l+1}\}}$. Since $f_{\boldsymbol{\tau}}|_{\mathcal{B}_l}$ and $f_{\boldsymbol{\tau}'}|_{\mathcal{B}_l}$ are linear and coincide in two points, we conclude that $f_{\boldsymbol{\tau}}|_{\mathcal{B}_l} = f_{\boldsymbol{\tau}'}|_{\mathcal{B}_l}$.

□

3.3 Constraints in the set $\mathcal{P}_C$ and the DNA replication context

In this section, we discuss how the constraints of the set $\mathcal{P}_C$ of Proposition 1 can be interpreted from the perspective of their application to the DNA analysis problem. In the context of DNA replication analysis, constant parts in vector $\boldsymbol{\tau}$ means that an interval of the DNA fragment $\mathcal{X}$ was replicated simultaneously, which is not in the physical hypothesis of the real problem. On the other hand, the spacing between break-points in the context of the application corresponds to the distance between two initiation to termination events. If these two events are very close, it is not possible for biologists to extract information from the BrdU signal, which makes this hypothesis reasonable. The spacing suggested by Proposition 1 is 12 space units, which corresponds to 1.2 kb in the actual signal. The non negativeness proposed in (1) also arises from the applied intuition to the problem.

4 DNA-inverse optimization

Existing numerical methods for nonlinear inverse problems provide local solutions to problem (**P1**) [5, 24]. However this appraoch faces two major limitations: (i) They are not able to provide a global solution, which is a major challenge in the DNA-inverse model; and (ii) They are not fast enough

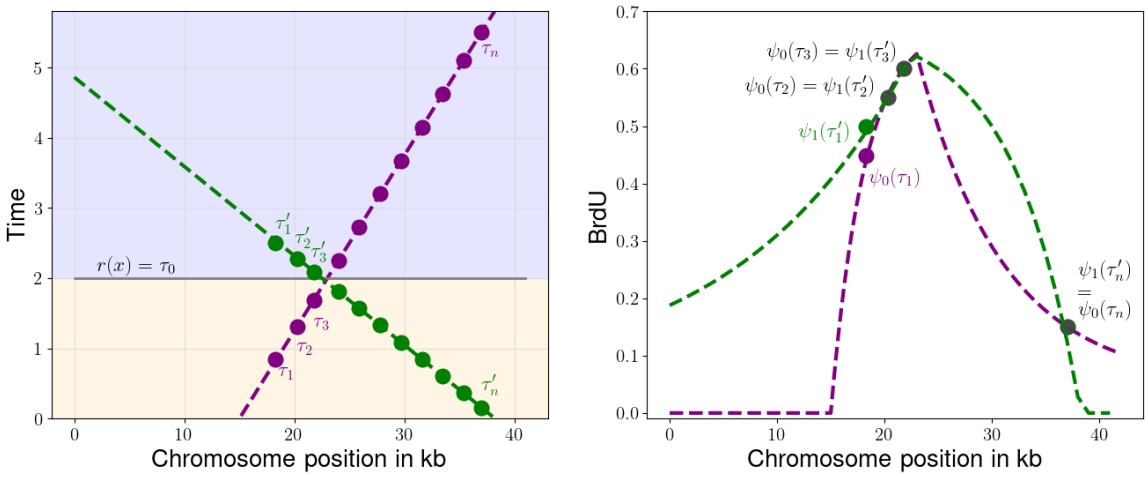

Figure 3: Illustration of the injectivity of Ψ for $C = 0$. (left) Purple dots represent the vector $\boldsymbol{\tau} = (\tau_1, \tau_2, \tau_3, ..., \tau_n)$ and green dots $\boldsymbol{\tau}' = (\tau_1', \tau_2', \tau_3', ..., \tau_n')$. Note that $\boldsymbol{\tau}, \boldsymbol{\tau}' \in \mathcal{P}_0$. The dashed lines represent the linear functions that support vectors $\boldsymbol{\tau}, \boldsymbol{\tau}'$ as stated in Lemma 1. We note that ψ is injective for vector components in the blue and orange regions. The main challenge in verifying Ψ injectivity is then to consider vectors $\boldsymbol{\tau}, \boldsymbol{\tau}'$ with components in different sides of the line $r(x) = \tau_0$, as those in the figure. (right) Dashed lines represent the images by Ψ of the lines that support vectors $\boldsymbol{\tau}$ (purple) and $\boldsymbol{\tau}'$ (green). The images coincide at only three coordinates: $1, 2$ and n. However, $\boldsymbol{\tau}$ and $\boldsymbol{\tau}'$ have different images. For example, in the case of component 3 : $\Psi(\boldsymbol{\tau})_3 = \psi_0(\tau_3) \neq \psi_1(\tau_3) = \Psi(\boldsymbol{\tau}')_3$. This case illustrates the proof of Lemma 2. In this lemma, we show that the set $\mathcal{I}_{+-}$, that consists on indexes for which components of τ, τ' are in different sides of the line $r(x) = \tau_0$, can not have three elements with coincident images.

to allow the exploration of different starting points. In this section, we reformulate problem (**P1**) in order to provide a numerical method able to achieve global solutions .

Employing the notation of Assumption **(A.2)**, the inverse set $\psi^{-1}(b)$, for any $b \in \mathbb{R}_+$, can be written as:

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

where $\psi_0^{-1}(b)$ and $\psi_1^{-1}(b)$ are inverse sets of ψ_0 and ψ_1 respectively. Because of (**A.2**), these sets consist in single elements or are empty. Consider a signal $\mathbf{z} \in \mathbb{R}^n$. Each $\boldsymbol{d} \in \{0,1\}^n$ is associated to a part of the inverse set $\Psi^{-1}(\mathbf{z})$, given by:

$$\Psi^{-1}(\mathbf{z}; \boldsymbol{d}) := \psi_{d_0}^{-1}(\mathrm{z}_0) \times ... \times \psi_{d_n}^{-1}(\mathrm{z}_n) \subset \mathbb{R}^n,$$

where some of the inverse sets on the right hand side can be empty. The set $\mathcal{K}(\mathbf{z}) \subset \{0,1\}^n$ select elements for which all inverse sets are composed by a single element:

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

For elements $\boldsymbol{d} \in \mathcal{K}$, $\Psi^{-1}(\mathbf{z}; \boldsymbol{d})$ can be identified as a vector in $\mathbb{R}^n$. Using this abuse of notation, we denote $\Psi_{\boldsymbol{d}}^{-1}(\mathbf{z})$ a vector in $\mathbb{R}^n$ when $\boldsymbol{d} \in \mathcal{K}$:

$$\boldsymbol{d} = (d_0, ..., d_n) \in \mathcal{K}(\mathrm{z}) \Rightarrow \Psi_{\boldsymbol{d}}^{-1}(\mathbf{z}) := (\psi_{d_1}^{-1}(\mathrm{z}_1), ..., \psi_{d_n}^{-1}(\mathrm{z}_n)) \in \mathbb{R}^n. \tag{5}$$

In this case, it is possible to develop a *Taylor expansion* of the function ψ, which is also a coordinatewise expansion of the function Ψ.*

Taylor expansion Consider a signal $\mathbf{z} \in \mathbb{R}^n$, and any coordinate $i \in \{1, ..., n\}$. Let $\boldsymbol{d} \in \mathcal{K}(\mathbf{z}) \subset \{0,1\}^n$, and $\boldsymbol{\tau} \in \mathbb{R}^n_+$. Note that for all $i \in \{1, ..., n\}$: $z_i = \psi_{d_i}(\psi_{d_i}^{-1}(z_i)) = \psi(\psi_{d_i}^{-1}(z_i))$. Then, if ψ is continuously differentiable, by Taylor's theorem applied in the point $\psi_{d_i}^{-1}(\mathbf{z}_i)$, for all $i \in \{1, ..., n\}$, we have:

$$\mathrm{z}_i - \psi(\tau_i) = \psi^{'}(\psi_{d_i}^{-1}(\mathrm{z}_i))\left(\psi_{d_i}^{-1}(\mathrm{z}_i) - \tau_i\right) + h(\tau_i)\left(\psi_{d_i}^{-1}(\mathrm{z}_i) - \tau_i\right)$$

where $\lim_{\tau_i \mapsto \psi_{d_i}^{-1}(\mathrm{z}_i)} h(\tau_i) = 0$. Define: $w_{\boldsymbol{d},i} := \psi^{'}(\psi_{d_i}^{-1}(\mathrm{z}_i)) \in \mathbb{R}$, for all $i \in \{1, ..., n\}$. Then, if $|\psi_{d_i}^{-1}(\mathrm{z}_i) - \tau_i|$ is small for all $i \in \{1, ..., n\}$:

$$\|\mathbf{z} - \Psi(\boldsymbol{\tau})\|_2^2 \approx \sum_{i=1}^{n} w_{\boldsymbol{d},i}^2 \left(\Psi_{\boldsymbol{d}}^{-1}(\mathbf{z}) - \boldsymbol{\tau}\right)_i^2. \tag{6}$$

The right hand side of (6) is a distance between $\Psi_{\boldsymbol{d}}^{-1}(\mathbf{z})$ and τ, that depends on $\boldsymbol{d}$. The variable $\boldsymbol{d}$ adjusts this distance to the local behavior of Ψ. To formalize this intuition, we define a pseudo-norm that coincide with this notion of distance for $\boldsymbol{d} \in \mathcal{K}(\mathbf{z})$.

Definition 2 (weighted norm)**.** *Given a vector $\boldsymbol{w} \in \mathbb{R}^n$ and any $\boldsymbol{v} \in \mathbb{R}^n$, the weighted norm $\|\boldsymbol{v}\|_{\boldsymbol{w}}$ with respect to the vector $\boldsymbol{w}$ is defined as:*

$$\|\boldsymbol{v}\|_{\boldsymbol{w}} := \sqrt{\sum_{i=1}^{n} w_i^2 v_i^2}.$$

Definition 3 ($\|\cdot\|_{\boldsymbol{w_d}}$)**.** *Let $\mathbf{z} \in \mathbb{R}^n_+$, and $\boldsymbol{d} \in \{0,1\}^n$. For any $\boldsymbol{v} \in \mathbb{R}^n$, we define: $\|\boldsymbol{v}\|_{\boldsymbol{w_d}}$, the weighted norm with respect to the vector $\boldsymbol{w_d}$, where:*

$$w_{\boldsymbol{d},i} := d_i \odot w_{\mathbf{0},i} + (1 - d_i) \odot w_{\mathbf{1},i}, \text{ for all } i \in \{1, ..., n\}, \tag{7}$$

and

$$w_{\mathbf{0},i} := \begin{cases} \psi^{'}(\psi_0^{-1}(\mathrm{z}_i)) & \psi_0^{-1}(\mathrm{z}_i) \neq \varnothing \\ 0 & \psi_0^{-1}(\mathrm{z}_i) = \varnothing, \end{cases} \qquad w_{\mathbf{1},i} := \begin{cases} \psi^{'}(\psi_1^{-1}(\mathrm{z}_i)) & \psi_1^{-1}(\mathrm{z}_i) \neq \varnothing \\ 0 & \psi_1^{-1}(\mathrm{z}_i) = \varnothing. \end{cases}$$

The definition 3 is illustrated in Figure 6. Note that when $\boldsymbol{d} \in \mathcal{K}(\mathbf{z})$, the definition of $\boldsymbol{w_d}$ coincides with the weights of the Taylor approximation (6). In Section 4.2, we discuss this definition to other values of $\boldsymbol{d} \in \{0,1\}^n$.

4.1 Noiseless signal

The Taylor expansion suggests that Problem (**P1**), defined using the Euclidean metric in $\mathbb{R}^n$, can be approximated by a problem that employs an alternative notion of distance given by $\|\cdot\|_{\boldsymbol{w_d}}$.

Alternative optimization problem in the noiseless case: Consider $\mathbf{z} \in \mathbb{R}^n_+$, and the following optimization problem:

$$(\boldsymbol{\tau}^*, \boldsymbol{d}^*) := \underset{\{(\boldsymbol{\tau},\boldsymbol{d}) \in \mathcal{P}_C \times \mathcal{K}(\mathbf{z})\}}{\arg\min} \|\boldsymbol{\tau} - \Psi_{\boldsymbol{d}}^{-1}(\mathbf{z})\|_{w_{\boldsymbol{d}}}^2, \tag{**P2**}$$

where $\|\cdot\|_{w_{\boldsymbol{d}}}$ is defined in Definition 3 and $\Psi_{\boldsymbol{d}}^{-1}(\mathbf{z})$ is defined as in (5).

In Figure 4, we observe possible different positions of the groundtruth timing profile $\bar{\boldsymbol{\tau}}$, and the solutions $(\boldsymbol{\tau}^*, d^*)$ of (**P2**) and $\widehat{\boldsymbol{\tau}}$ of (**P1**).

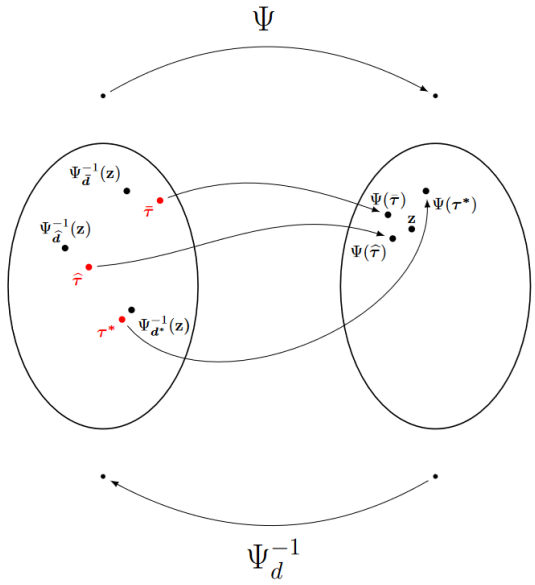

Figure 4: Illustration of formulation (**P1**) and (**P2**). Elements in red are in the set $\mathcal{P}_C$, the set of piecewise linear vectors with at most C breakpoints. On the right, we observe the set where Problem (**P1**) is formulated, with the Euclidean distance defining the fidelity term. With formulation (**P2**'), we transfer the optimization problem to the domain of Ψ using an auxiliary integer variable $\boldsymbol{d} \in \{0,1\}^n$. Depending on $\boldsymbol{d}$, the inverse $\Psi_{\boldsymbol{d}}^{-1}(\mathbf{z})$ is situated in different parts of the domain, justifying the introduction of $\|\cdot\|_{w_{\boldsymbol{d}}}$ to account for the local behavior of Ψ

Proposition 2 (Noiseless case)**.** *Given any* $\bar{\boldsymbol{\tau}} \in \mathcal{P}_C^{\geqslant}$ *defined in* (3)*, consider* $\mathbf{z} = \Psi(\bar{\boldsymbol{\tau}})$*. Then* (**P1**) *and* (**P2**) *both admit the same unique solution in* $\mathcal{P}_C^{\geqslant}$*, which is precisely* $\boldsymbol{\tau}$*.*

Proof. Clearly $\widehat{\boldsymbol{\tau}} = \bar{\boldsymbol{\tau}}$ is a solution for **(P1)**. On the other hand, consider $\bar{\boldsymbol{d}} \in \{0,1\}^n$ defined as: $\bar{d}_i = 0$, if $\bar{\tau}_i \leqslant \tau_0$ and $\bar{d}_i = 1$, if $\bar{\tau}_i > \tau_0$, for all $i \in \{1,...,n\}$. Then $\bar{\boldsymbol{\tau}} = \Psi_{\bar{\boldsymbol{d}}}^{-1}(\mathbf{z})$, and $\boldsymbol{d} \in \mathcal{K}(\mathbf{z})$, implying that $(\bar{\boldsymbol{\tau}}, \bar{\boldsymbol{d}})$ is a solution for problem (**P2**). Because of Proposition 1, if $\bar{\boldsymbol{\tau}} \in \mathcal{P}_C^{\geqslant}$, this solution is unique on this set. □

4.2 Noisy signal

In the case of a noisy signal $\mathbf{z} \in \mathbb{R}^n$, there is no guarantee that the optimal solution $\boldsymbol{d}^*$ of (**P2**) will be in the set $\mathcal{K}(\mathbf{z})$. To see that, write $\mathbf{z} = \Psi(\bar{\boldsymbol{\tau}}) + \epsilon$, where $\bar{\boldsymbol{\tau}} \in \mathcal{P}_C$ and ϵ is a random vector of dimension n. Let $\bar{\mathbf{z}} = \Psi(\bar{\boldsymbol{\tau}})$. If ϵ is small enough, we expect the solution of problem (**P2**) to be $(\boldsymbol{\tau}^*, \boldsymbol{d}^*)$, where $\boldsymbol{\tau}^* = \bar{\boldsymbol{\tau}}$, and $\boldsymbol{d}^*$ is such that $\bar{\boldsymbol{\tau}} = \boldsymbol{\tau}^* = \Psi_{\boldsymbol{d}^*}^{-1}(\bar{\mathbf{z}})$. Clearly $\boldsymbol{d}^* \in \mathcal{K}(\bar{\mathbf{z}})$. On the other hand, there is no reason for $\boldsymbol{d}^* \in \mathcal{K}(\mathbf{z})$. To extend Problem (**P2**) to a noisy signal, we need to define the vector $\Psi_{\boldsymbol{d}}^{-1}(\mathbf{z})$ when $\boldsymbol{d} \notin \mathcal{K}(\mathbf{z})$. In this extension, we take into account indices $i \in \{1,...,n\}$ for which $\psi^{-1}(\mathrm{z}_i) = \varnothing$.

Definition 4 ($\Psi_{\boldsymbol{d}}^{-1}(\mathbf{z})$ and $\|\cdot\|_{\boldsymbol{w_d}}$)**.** *Let* $\mathbf{z} \in \mathbb{R}_+^n$*, and* $\boldsymbol{d} \in \{0,1\}^n$*. We extend the definition of* $\Psi_{\boldsymbol{d}}^{-1}(\mathbf{z})$

for all $\boldsymbol{d} \in \{0,1\}^n$. *For* $i \in \{1,...,n\}$*:*

$$\Psi_{\boldsymbol{d}}^{-1}(\mathbf{z})_i := \begin{cases} \psi_{d_i}^{-1}(\mathbf{z}_i), & \text{If } \psi_{d_i}^{-1}(\mathbf{z}_i) \neq \varnothing \\ \infty, & \text{If } \psi_{d_i}^{-1}(\mathbf{z}_i) = \varnothing, \end{cases}$$

where the notation $\psi_{d_i}^{-1}(\mathrm{z}_i)$ *is used both for the inverse set and inverse function. Denote* $\bar{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$*, and* $\boldsymbol{w_d}$ *defined in Definition 3. For any* $\boldsymbol{v} \in \bar{\mathbb{R}}^n$*, we define:*

$$\|\boldsymbol{v}\|_{\boldsymbol{w_d}} = \sqrt{\sum_{\{i \in [n]: \boldsymbol{v}_i \neq \infty\}} w_{\boldsymbol{d},i}^2 v_i^2}.$$

In practice, Definition 4 extends the definition of $\|\cdot\|_{\boldsymbol{wd}}$ to values where $\boldsymbol{d} \notin \mathcal{K}(\mathbf{z})$, but considers only indices i where $\psi_{d_i}^{-1}(z_i)$ is a singleton. Two reasons can motivate this definition. First, in the noiseless case, weights extend continuously to zero in the region where ψ is not surjective. Observing Figure 1 C, and defining $a := \lim_{t\to\infty} \psi_1$, we note that ψ_1 becomes indefinitely close to a constant function as t tends to infinity, causing the corresponding weight $\boldsymbol{w}_{1,i}$ to tend to zero for indices $i \in \{1,...,n\}$ where z_i is close, but above, a. In this context, a natural extension to the weights for indices $i \in \{1,...,n\}$ where z_i is below a is zero. In the noisy case, $\boldsymbol{d} \notin \mathcal{K}(\mathbf{z})$ can also be the effect of noise, for example, in regions where $\mathbf{z}$ is above the maximum $\psi_{\max} := \max_{t\in[0,\infty)} \psi(t)$. In these cases, the choice of not taking into account these indices in $\|\cdot\|_{\boldsymbol{w_d}}$ is motivated by the numerical results presented in Section 6.

Alternative optimization problem for the noisy case: The extension of problem (**P2**) for a noisy signal $\mathbf{z}$ reads:

$$(\boldsymbol{\tau}^*, \boldsymbol{d}^*) := \underset{\{(\boldsymbol{\tau},\boldsymbol{d}) \in \mathcal{P}_C \times \{0,1\}^n\}}{\arg\min} \|\boldsymbol{\tau} - \Psi_{\boldsymbol{d}}^{-1}(\mathbf{z})\|_{w_{\boldsymbol{d}}}^2, \tag{**P2**'}$$

where $\|\cdot\|_{w_{\boldsymbol{d}}}$ and $\Psi_{\boldsymbol{d}}^{-1}(\mathbf{z})$ are defined in Definition 4.

5 Numerical approach

Note that the *mixed integer nonlinear problem* (**P2**') can be reformulated as follows:

$$\begin{array}{ll} \min\limits_{\boldsymbol{\tau},\boldsymbol{d}} & \frac{1}{2}\|\boldsymbol{d} \odot (\boldsymbol{\tau} - \mathbf{z^1})\|_{\boldsymbol{w_1}}^2 + \frac{1}{2}\|(\mathbf{1} - \boldsymbol{d}) \odot (\boldsymbol{\tau} - \mathbf{z^0})\|_{\boldsymbol{w_0}}^2 \\ \text{s.t.} & \boldsymbol{\tau} \in \mathbb{R}^n,\ \|\mathrm{L}\boldsymbol{\tau}\|_0 \leqslant C \\ & \boldsymbol{d} \in \{0,1\}^n, \end{array} \tag{8}$$

where the operator $\odot$ represents the coordinatewise multiplication. $\mathbf{1}, \mathbf{0} \in \{0,1\}^n$ are constant vectors, $\mathbf{z^1} := \Psi_{\mathbf{1}}^{-1}(\mathbf{z})$ and $\mathbf{z^0} := \Psi_0^{-1}(\mathbf{z})$. $\boldsymbol{w_1}$ and $\boldsymbol{w_0}$, are defined in (7).

The advantages of this reformulation are twofold: First, for each fixed $\boldsymbol{d}$, problem (8) has a quadratic objective function with a non-convex constraint. This optimization problem can be approximately solved by ℓ_1 regularization, in a formulation similar to *generalized lasso* [22]. On the other hand, even if the available set in (8) contains a large set of integer variables, we can attempt to reduce this set based on observations about the nature of solutions of problem (**P2**'). This analysis will be developed in Section 5.2.

5.1 Combinatorial method for DNA-Inverse

In this section, we present a methodology to address problem (8). Our approach involves solving this problem iteratively for each fixed $\boldsymbol{d}$, while comparing the obtained optimal values. Clearly, directly applying this method to the entire set $\{0,1\}^n$ would be not tractable due to its exponential size. To mitigate this problem, we introduce a subset $\mathcal{D} \subset \{0,1\}^n$ in which the optimal solution $\boldsymbol{d}^* \in \mathcal{D}$. The specific computation of this subset will be discussed in Section 5.3.

For each $\boldsymbol{d} \in \mathcal{D}$, we propose to relax the optimization problem:

$$\begin{array}{ll} \min\limits_{\boldsymbol{\tau}} & \frac{1}{2}\|\boldsymbol{d} \odot (\boldsymbol{\tau} - \mathbf{z^1})\|^2_{\boldsymbol{w_1}} + \frac{1}{2}\|(\mathbf{1} - \boldsymbol{d}) \odot (\boldsymbol{\tau} - \mathbf{z^0})\|^2_{\boldsymbol{w_0}} \\ \text{s.t.} & \boldsymbol{\tau} \in \mathbb{R}^n,\ \|\mathrm{L}\boldsymbol{\tau}\|_0 \leqslant C \end{array} \tag{9}$$

into the l_1 regularized problem:

$$\boldsymbol{\tau}^*_{\boldsymbol{d}} := \operatorname*{argmin}_{\boldsymbol{\tau}} \ \frac{1}{2}\|\boldsymbol{d} \odot (\boldsymbol{\tau} - \mathbf{z^1})\|^2_{\boldsymbol{w_1}} + \frac{1}{2}\|(\mathrm{I} - \boldsymbol{d}) \odot (\boldsymbol{\tau} - \mathbf{z^0})\|^2_{\boldsymbol{w_0}} + \lambda\|\mathrm{L}\boldsymbol{\tau}\|_1 \tag{10}$$

The solutions $\boldsymbol{\tau}^*_{\boldsymbol{d}}$, for $\boldsymbol{d} \in \mathcal{D}$ will be compared using the objective function criterion (9):

$$\boldsymbol{\tau}^* := \min_{\{\boldsymbol{d} \in \mathcal{D}\}} F(\boldsymbol{\tau}^*_{\boldsymbol{d}}), \text{ where } F(\boldsymbol{\tau}^*_{\boldsymbol{d}}) := \frac{1}{2}\|\boldsymbol{d} \odot (\boldsymbol{\tau}^*_{\boldsymbol{d}} - \mathbf{z^1})\|^2_{\boldsymbol{w_1}} + \frac{1}{2}\|(\mathbf{1} - \boldsymbol{d}) \odot (\boldsymbol{\tau}^*_{\boldsymbol{d}} - \mathbf{z^0})\|^2_{\boldsymbol{w_0}} \tag{11}$$

Dual of Generalized-lasso: Note that problem (10) can be written as a *generalized-lasso* problem such as in [22, Equation (2)]:

$$\operatorname*{argmin}_{\boldsymbol{\tau}} \frac{1}{2}\|\boldsymbol{w} \odot (\boldsymbol{\tau} - \mathbf{z^d})\|^2_2 + \lambda\|\mathrm{L}\boldsymbol{\tau}\|_1 \tag{12}$$

where $\mathbf{z^d} = \boldsymbol{d} \odot \mathbf{z^1} + (1 - \boldsymbol{d}) \odot \mathbf{z^0}$, and $\boldsymbol{w} \in \mathbb{R}^n$ is defined by:

$$w_i = \begin{cases} d_i w_{\mathbf{1},i} + (1 - d_i) w_{\mathbf{0},i}, & \text{if } w_{\mathbf{1},i} \text{ or } w_{\mathbf{0},i} \neq 0 \\ 0, & \text{if } w_{\mathbf{1},i} = 0 \text{ and } w_{\mathbf{0},i} = 0. \end{cases}$$

In Theorem [22, Section 4], a quadratic expression for the dual of the generalized-lasso is presented for the case $\boldsymbol{w} = \mathbf{1}$. We aim to extend this result to the scenario where $\boldsymbol{w}$ is any vector. To facilitate this extension, we introduce the following notation:

$$\mathcal{I}^+ = \{i \ : \ w_i > 0\}, \ \ \mathcal{I}^0 = \{i \ : \ w_i = 0\}$$

Proposition 3. *Let $\boldsymbol{w}^{-1}$ defined by: $w_i^{-1} := \frac{1}{w_i}$, for all $i \in \mathcal{I}^+$ and $w_i^{-1} := 0$, for $i \in \mathcal{I}^0$. Let $\boldsymbol{u}^*$ be the solution of the following optimization problem:*

$$u^* = \begin{array}{lll} & \textit{argmin} & \frac{1}{2}\|\boldsymbol{w} \odot (\mathrm{L}^\top \boldsymbol{u})\|^2 - \langle \mathrm{L}\mathbf{z^d}, \boldsymbol{u} \rangle \\ & \textit{s.t.} & \|\boldsymbol{u}\|_\infty \leqslant \lambda \\ & & (\mathrm{L}^\top \boldsymbol{u})_i = 0, \ \textit{for } i \in \mathcal{I}^0 \end{array} \tag{13}$$

Then, $\boldsymbol{\tau}^*$ defined by:

$$\begin{cases} \tau_i^* = \mathbf{z^d} - (\boldsymbol{w}^2)^{-1} \odot (\mathrm{L}^\top \boldsymbol{u}^*), & i \in \mathcal{I}^+ \\ (\mathrm{L}\boldsymbol{\tau}^*)_i = 0 & i \in \mathcal{I}^0 \end{cases}$$

is a solution of problem (12).

Proof. Denote $f(\tau) = \frac{1}{2}\|\boldsymbol{w} \odot (\boldsymbol{\tau} - \mathbf{z}^d)\|_2^2 + \lambda\|\mathrm{L}\boldsymbol{\tau}\|_1$ the convex objective function of problem (12). A solution $\boldsymbol{\tau}^*$ must satisfy:

$$0 \in \partial f(\boldsymbol{\tau}^*).$$

Note that this equation can be written componentwisely and when $i \in \mathcal{I}^0$, this implies that:

$$(\mathrm{L}\boldsymbol{\tau}^*)_i = 0, \text{ for } i \in \mathcal{I}^0 \tag{14}$$

On the other hand, problem (12) can be expressed through its dual problem:

$$\begin{array}{l} \max\limits_{u} \min\limits_{\boldsymbol{\tau}} \quad \dfrac{1}{2}\|\boldsymbol{w} \odot (\boldsymbol{\tau} - \mathbf{z}^{\boldsymbol{d}})\|_2^2 + \langle \boldsymbol{\tau}, \mathrm{L}^\top \boldsymbol{u}\rangle \\ \text{s.t.} \quad \|\boldsymbol{u}\|_\infty \leqslant \lambda. \end{array} \tag{15}$$

The minimization in $\boldsymbol{\tau}$ has the optimal conditions:

$$\begin{array}{l} \tau_i = \mathbf{z}^{\boldsymbol{d}}{}_i - ((\boldsymbol{w}^2)^{-1} \odot (\mathrm{L}^\top \boldsymbol{u}))_i, \quad i \in \mathcal{I}^+ \\ Ł^\top \boldsymbol{u}_i = 0, \quad i \in \mathcal{I}^0 \end{array} \tag{16}$$

replacing these conditions in problem (15), we obtain the dual variable u^* as a solution of problem (13). Joining conditions in (14) and (16) we have:

$$\begin{cases} \tau_i^* = \mathbf{z}^{\boldsymbol{d}} - (\boldsymbol{w}^2)^{-1} \odot (\mathrm{L}^\top \boldsymbol{u}^*), \quad i \in \mathcal{I}^+ \\ (\mathrm{L}\boldsymbol{\tau}^*)_i = 0 \quad i \in \mathcal{I}^0 \end{cases}$$

□

5.2 Constraints of the set $\mathcal{D}$

In this section, we discuss how to reduce the set of integer variables in problem (8). User In principle, the solution $\boldsymbol{d}^*$ could be any element of $\{0,1\}^n$. A more detailed examination of the problem will show that additional constraints can be incorporated in the available set of (8) without changing its solution. This analysis will be conducted in two parts: In the first part, we will identify indices for which $\boldsymbol{d}$ is allowed to oscillate, that is, when $d_i = 0$ and $d_{i+1} = 1$ (or the opposite). In the second part, we will consider the values where the signal $\mathbf{z} = 0$ and its impact on the variable $\boldsymbol{d} \in \{0,1\}^n$. We aim defining a subset $\mathcal{D} \subset \{0,1\}^n$ of the form:

$$\mathcal{D} = \{\boldsymbol{d} : \mathrm{B}\boldsymbol{d} = \mathbf{0}_b, \ \mathrm{A}\boldsymbol{d} = \mathbf{0}_a\}, \tag{17}$$

where the linear constraints are defined by matrices $\mathrm{A} \in \mathbb{R}^{n,a}$, and $\mathrm{B} \in \mathbb{R}^{n,b}$, and $\mathbf{0}_a \in \mathbb{R}^a$ and $\mathbf{0}_b \in \mathbb{R}^b$ are constant vectors.

Oscillations on $\boldsymbol{d}$: Consider a noiseless signal $\mathbf{z} = \Psi(\bar{\boldsymbol{\tau}})$, such that $\bar{\boldsymbol{\tau}} \in \mathcal{P}_C$. Then there exists $\bar{\boldsymbol{d}} \in \{0,1\}^n$, such that $\bar{\boldsymbol{\tau}} = \Psi_{\bar{\boldsymbol{d}}}^{-1}(\mathbf{z})$. Suppose that $(\bar{\boldsymbol{\tau}}, \bar{\boldsymbol{d}})$ is unknown. We aim to define the smallest possible set $\mathcal{D}$ such that $\bar{\boldsymbol{d}} \in \mathcal{D}$.

For $\bar{\boldsymbol{\tau}} \in \mathcal{P}_C$, the number of indices for which $\bar{\boldsymbol{\tau}}$ crosses τ_0 is bounded by C :

$$\#\{i : \bar{\tau}_i \leqslant \tau_0 \leqslant \bar{\tau}_{i+1} \ \text{ or } \bar{\tau}_{i+1} \leqslant \tau_0 \leqslant \bar{\tau}_i\} \leqslant C.$$

On the other hand, for $i \in \{1, ..., n\}$:

$$\bar{\tau}_i = \Psi_{\bar{\boldsymbol{d}}}^{-1}(\mathbf{z})_i = \begin{cases} \psi_{\mathbf{0}}^{-1}(\mathrm{z}_i) \in [0, \tau_0], \text{ if } \bar{d}_i = 0 \\ \psi_{\mathbf{1}}^{-1}(\mathrm{z}_i) \in [\tau_0, \infty), \text{ if } \bar{d}_i = 1. \end{cases}$$

Then, indices where $\bar{\boldsymbol{d}}$ change between 0 and 1 and indices where $\bar{\boldsymbol{\tau}}$ passes through τ_0 are the same, as illustrated in Figure 5. For example:

$$\bar{d}_i = 0 \text{ and } \bar{d}_{i+1} = 1 \Rightarrow \bar{\tau}_i \leqslant \tau_0 \leqslant \bar{\tau}_{i+1}.$$

Then, we define the set of transitions on $\boldsymbol{d}$ by:

$$\mathcal{I}^A := \{i : \bar{d}_i + \bar{d}_{i+1} = 1\} = \{i : \bar{\tau}_i \leqslant \tau_0 \leqslant \bar{\tau}_{i+1} \text{ or } \bar{\tau}_{i+1} \leqslant \tau_0 \leqslant \bar{\tau}_i\}.$$

From this characterization, we aim computing $\mathcal{I}^A$ using only the known signal $\mathbf{z}$ and the function Ψ. To achieve this, we draw inspiration from the case where the signal $\mathbf{z}$ is defined continuously. Let $z : I \subset \mathbb{R} \to \mathbb{R}$ be a continuous function, where I is an interval. Consider the functions $\psi_1^{-1}(z) : I \to [\tau_0, \infty)$, $\psi_0^{-1}(z) : I \to [0, \tau_0]$ and $\bar{d} : I \to \{0, 1\}$. Let $\bar{\tau}$ be defined as:

$$\bar{\tau}(x) = \begin{cases} \psi_0^{-1}(z(x)) \in [0, \tau_0], \text{ if } \bar{d}(x) = 0 \\ \psi_1^{-1}(z(x)) \in [\tau_0, \infty), \text{ if } \bar{d}(x) = 1. \end{cases}$$

Denote $f := \psi_1^{-1}(z) - \psi_0^{-1}(z)$. Then, if $\bar{\tau}$ crosses τ_0, it exists $\bar{x} \in I$ such that $\tau(\bar{x}) = \tau_0$, implying $f(\bar{x}) = 0$. Since $f \geqslant 0$, $\bar{x}$ is a local minima of f. The reasoning is illustrated in Figure 5. We will employ this continuous intuition in a vectorial context, i.e, detect indices in $\mathcal{I}^A$ as local minima of the vector:

$$\Psi_{\mathbf{0}}^{-1}(\mathbf{z}) - \Psi_{\mathbf{1}}^{-1}(\mathbf{z}).$$

note that only the input signal $\mathbf{z}$ is necessary to compute this vector. Then:

$$\mathcal{D} = \{\boldsymbol{d} \in \{0, 1\}^n, d_i = d_{i+1} \text{ for } i \in \mathcal{I}^A\}$$

This strategy is interesting since we can extract *a priori* information about the indices where the solution $\bar{\boldsymbol{d}}$ can oscillate between 0 and 1. Considering that we are able to detect $\mathcal{I}^A = \{i_1, ..., i_J\}$, we define a matrix $\mathrm{A} \in \mathbb{R}^{n \times a}$ in such a way that for $j \in \{1, ..., J\}$:

$$\begin{aligned} &\mathrm{A}[i, j] = 1, \text{ for } i = i_j \\ &\mathrm{A}[i, j] = -1, \text{ for } i = i_j + 1 \\ &\mathrm{A}[i, j] = 0, \text{ for } i \neq i_j, i_j + 1. \end{aligned}$$

If $\boldsymbol{d} \in \{0, 1\}^n$ is a solution of problem (8), it implies that $\mathrm{A}\boldsymbol{d} = \mathbf{0}_m \in \mathbb{R}^m$.

Zeros of z : Another source of information are indices where the signal $\mathbf{z} = 0$. Note that for any $i \in \{1, ..., n\}$ $\mathrm{z}_i = 0$ implies $\tau_i = 0$, which implies $\boldsymbol{d} = 0$. Then the set: $\mathcal{I}^B = \{i : \mathrm{z}_i = 0\}$, is also important. The set $\mathcal{D}$ is defined by:

$$\mathcal{D} = \{\boldsymbol{d} \in \{0, 1\}^n : d_i = d_{i+1} \text{ for } i \in \mathcal{I}^A, \ d_i = 0 \text{ for } i \in \mathcal{I}^B\}, \tag{18}$$

The matrix B has as columns canonical vectors for indices in $\mathcal{I}^B$. The set $\mathcal{D}$ can be written as in (17). In Section 5.3, we discuss how the noise in $\mathbf{z}$ impacts the definition of $\mathcal{D}$ for real data.

5.3 Algorithm

In this section, we discuss the algorithm that provides a numerical solution for Problem (**P2**'). The initialization of the algorithm accounts for the high presence of noise in the real DNA replication data, and occurs in two steps: 1. The computation of the weights $\boldsymbol{w_d}$ defined (7); 2. The computation of the set $\mathcal{D}$ defined in (18). The algorithm is described in Algorithm 1.

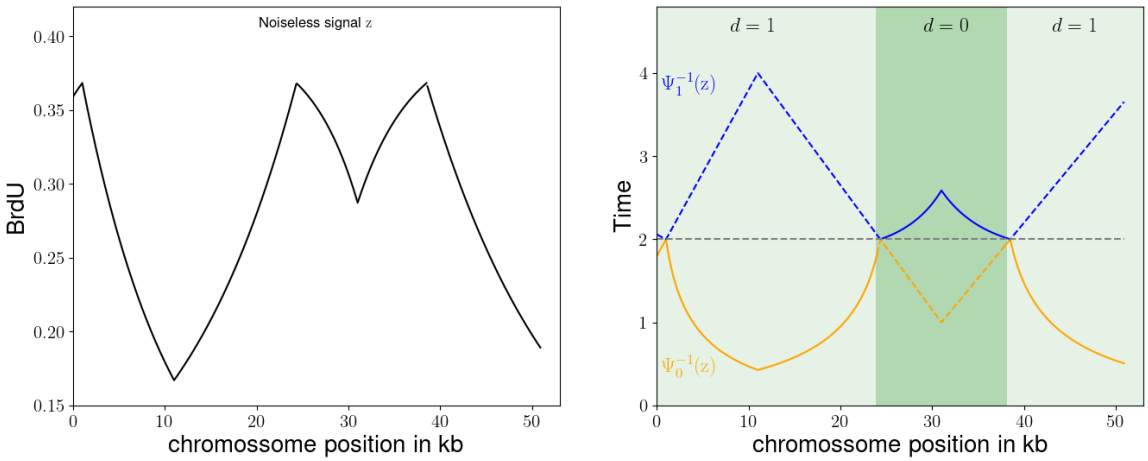

Figure 5: Illustration of possible oscillations of $\bar{\boldsymbol{d}} \in \mathcal{D}$. (left) Noiseless signal $\mathbf{z} = \Psi(\bar{\boldsymbol{\tau}})$ for $\bar{\boldsymbol{\tau}} \in \mathcal{P}_C$. (right) Inverses $\Psi_{\mathbf{1}}^{-1}(\mathbf{z})$ (blue) and $\Psi_{\mathbf{0}}^{-1}(\mathbf{z})$ (orange). The piecewise linear $\bar{\boldsymbol{\tau}}$ can be observed by the dashed lines. For each index $i \in \{1, ..., n\}$, when $\bar{\tau}_i$ is blue, $\bar{d}_i = 1$. When $\bar{\tau}_i$ is orange, $\bar{d}_i = 0$. When d oscillates, τ crosses the gray dashed line $r(x) = \tau_0 = 2$. Note the oscillations of $\bar{\boldsymbol{d}}$, illustrated by transitions in shades of green, coincide with indices where $\Psi_{\mathbf{1}}^{-1}(\mathbf{z}) = \Psi_{\mathbf{0}}^{-1}(\mathbf{z})$. In addition: $\Psi_{\mathbf{1}}^{-1}(\mathbf{z}) - \Psi_{\mathbf{0}}^{-1}(\mathbf{z}) \geqslant 0$. We conclude that oscilations in $\boldsymbol{d}$ are local minima of the vector $\Psi_{\mathbf{1}}^{-1}(\mathbf{z}) - \Psi_{\mathbf{0}}^{-1}(\mathbf{z})$.

Weights: The weights $\boldsymbol{w_d}$ are computed from weights $w_{\mathbf{0}}$ and $w_{\mathbf{1}}$, for $\mathbf{0}, \mathbf{1} \in \mathbb{R}^n$ in the following way: $\boldsymbol{w_d} = \boldsymbol{d} \odot \boldsymbol{w_1} + (1 - \boldsymbol{d}) \odot \boldsymbol{w_0}$. The calculus of weights $\boldsymbol{w_0}, \boldsymbol{w_1}$ is done following Definition 3. For stability reasons, a smoother version $\widetilde{\mathbf{z}}$ of the signal $\mathbf{z}$ replaces the original signal in this calculus. This computation uses the Savitzky-Golay Smoothing Filter that results in $\widetilde{\mathbf{z}}$ such that for each $i \in \{1, ..., n\}$: $\widetilde{\mathrm{z}}_i = \sum_{i=-2}^{2} c_n \mathrm{z}_i$, where $c_n = 1/5$. The weights $\boldsymbol{w_0}, \boldsymbol{w}_1$ are illustrated in Figure 6 for a noisy signal $\mathbf{z}$.

Set $\mathcal{D}$: The set $\mathcal{D}$, defined in (18) depends on the set of indices $\mathcal{I}^B$ and $\mathcal{I}^A$. For the noiseless case, the indices in $\mathcal{I}^B$ are the indices where the signal $\mathbf{z}$ is equal to zero. In the noisy case, this detection stills simple, since the zeros of the signal $\mathbf{z}$ are subjected to considerably less noise then the rest of the signal (see Figure 2). Then:

$$\mathcal{I}^B = \{i : \mathrm{z}_i = 0 \text{ and } \mathrm{z}_{i+1} = 0\}.$$

and $b := \#\mathcal{I}^B$. In the case of the set $\mathcal{I}^A$, as discussed in Section 5.2, the computation involves the vector:

$$\boldsymbol{h} := \Psi_{\mathbf{0}}^{-1}(\mathbf{z}) - \Psi_{\mathbf{1}}^{-1}(\mathbf{z}). \tag{19}$$

The indices of local minima of h indicate possible oscillations on $\boldsymbol{d}$ (from 0 to 1 or the opposite). In python, many packages are available to compute this local minima. Since this vector have the same size of the signal $\mathbf{z}$, which ranges from 100 to 1000, the time required for this computation is negligible. Due to noise, the local minima of $\boldsymbol{h}$ are treated as the centers of intervals with a certain size: s^A (an input parameter) in which $\boldsymbol{d}$ is allowed to oscillate. Denote $\mathcal{M} = \{i_1^{\boldsymbol{h}}, ..., i_k^{\boldsymbol{h}}\}$ the indices of local minimima of $\boldsymbol{h}$. Denote $I_1^{\boldsymbol{h}}, ..., I_k^{\boldsymbol{h}}$, the correspondent intervals of size s^A centered on the correspondent elements of $\mathcal{M}$. Then the set $\mathcal{I}^A$ is defined as:

$$\mathcal{I}^A = \{i \in \{1, ..., n\} : i \notin I_j^{\boldsymbol{h}} \text{ for } j \in \{1, ..., k\}\}.$$

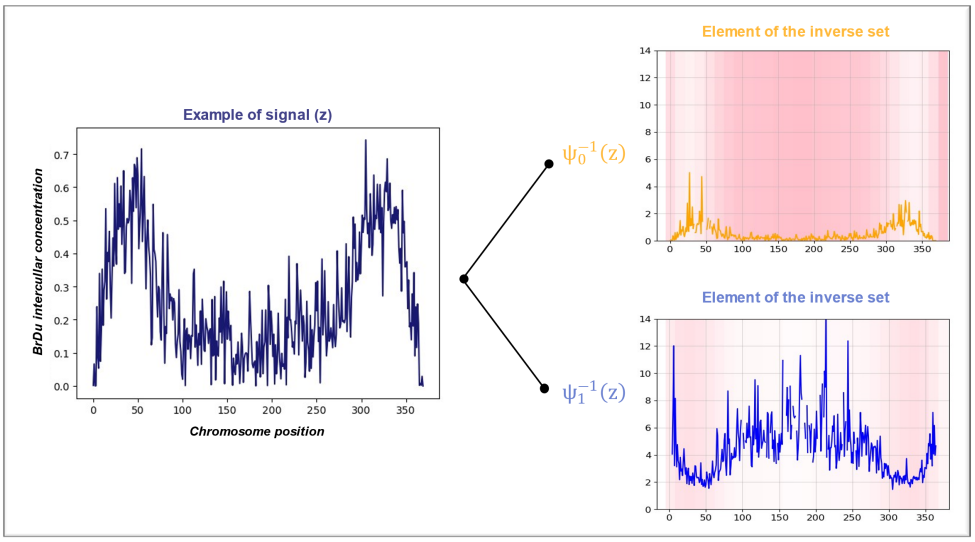

Figure 6: Illustration of $\Psi_{\boldsymbol{d}}^{-1}(\mathbf{z})$ and $\boldsymbol{w_d}$ in the noisy case, for $\boldsymbol{d} = \mathbf{0}, \mathbf{1} \in \mathbb{R}^n$. (left) Noisy read. (right) Vectors $\Psi_{\mathbf{1}}^{-1}(\mathbf{z})$ and $\Psi_{\mathbf{0}}^{-1}(\mathbf{z})$. The vector $\Psi_{\mathbf{1}}^{-1}(\mathbf{z})$ is not represented in indices where $\Psi_{\mathbf{1}}^{-1}(\mathbf{z}) = \infty$. We observe that the local behavior of Ψ expands or contracts the noise present in the signal. For this reason, to compare the precision of a piecewise linear approximation of these curves, it is necessary to define $\|\cdot\|_{\boldsymbol{w_d}}$. The weights in $\boldsymbol{w_0}, \boldsymbol{w_1}$ are represented by the intensity of the pink background for $\Psi_{\mathbf{1}}^{-1}(\mathbf{z})$ and $\Psi_{\mathbf{0}}^{-1}(\mathbf{z})$ respectively. The more intense the background, the greater the value of $\boldsymbol{w_0}, \boldsymbol{w_1}$ associated with this part of the signal. We note that more significant noise requires smaller weights.

Then $a := n - ks^A$. Clearly, the set $\mathcal{D}$ is capable of drastically reducing the possible solutions $\boldsymbol{d} \in \{0,1\}^n$. Nevertheless, the best upper bound for $\mathcal{D}$ cardinality is: $\#\mathcal{D} \leqslant (s^A)^k$ which stills large. Therefore, we consider a representative subset of $\tilde{\mathcal{D}} \subset \mathcal{D}$. It work as follows: we divide each interval $I_i^{\boldsymbol{h}}$ into m^A (an input parameter) equal parts, and consider the possible combinations for $i \in \{1, ..., k\}$. Thus, the set $\#\tilde{\mathcal{D}} \leqslant (m^A)^k$ elements. In general, for DNA replication signals considered in the application, k varies between 2 and 5.

Choice of parameters: The parameter s^A is choose as 60 units, or $0.6kb$. The parameter m is chose as 3. These choices aim to balance good results with low computation time. Higher values of m directly impact the computation time as they increase the number of elements in $\tilde{\mathcal{D}}$. The value of s^A is more associated with the amount of noise and uncertainty regarding the local minima of the vector h in (19). The parameter λ is empirically fixed as 8.

Algorithm: According to Section 5.1, the DNA-inverse algorithm loops as follows:

6 Numerical results

The numerical results presented in this section utilize real data obtained from yeast DNA [14]. We opt for real data due to the challenge of fully replicating the type of noise present in this dataset. The signals vary in size from 50 to 1000, and a total of 300 reads are analyzed with an average

Algorithm 1: DNA-Inverse

Data: Input data: $\mathbf{z}$. Parameters: s^A and m^A.
Initialization:
Compute weights $w_{\boldsymbol{d}}$. Compute the sets $\tilde{\mathcal{D}} \subset \mathcal{D}$. Set: $\mathcal{D}_{\text{past}} = \varnothing$;
Main Loop:
for $\boldsymbol{d} \in \tilde{\mathcal{D}}$ **do**
  Step 0: $\mathcal{D}_{\text{past}} \leftarrow \mathcal{D}_{\text{past}} \cup \{\boldsymbol{d}\}$;
  Step 1: Solve the optimization problem (10) by its dual formulation (13) (quadratic optimization), obtaining a solution $\boldsymbol{\tau}^*_{\boldsymbol{d}}$;
  Step 2: Compute the objective of problem (8) for $\boldsymbol{\tau}^*_{\boldsymbol{d}}$ (without ℓ_1 regularization) and compare with objective values of previous $\boldsymbol{d} \in \mathcal{D}_{\text{past}}$:

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

  F is defined in (11).
Output: τ^*, d^*

of 2.3 detected forks by read. Biologists, with their trained eye, are capable of recognizing the replication events associated with each of these reads. The objective of this test is to detect these events automatically. The FORQ-seq technique [13] has enabled scientists to significantly increase the amount of data available for analysis. The challenge now is to develop methods that do not rely on individual analysis of each sample, allowing for a considerable increase in the amount of analyzed data.

The numerical results from the DNA-Inverse method will be divided into two parts: Section 6.1 compares DNA-Inverse with an state-of-the-art proximal method capable of providing local minima to (**P1**). In Section 6.2, we explore the advantages of this method and discuss its relevance for DNA replication analysis.

6.1 Comparison with state-of-the-art method

When Ψ is non-linear, the resulting optimization problem (**P1**) is generally non-convex, which is a major challenge in optimization. Recent theoretical and numerical advancements have significantly improved the treatment of problems of type (**P1**), showing remarkable flexibility concerning the types of operators Ψ and regularization terms [19, 5, 25, 9]. However, a key limitation of these methods lies in their pursuit of local solutions, due to the non-convex nature of the problem. In relevant applications, such as DNA replication analysis, attaining only local solutions do not provide substantial progress towards achieving the overall objective. In these cases, numerical methods should focus on special applications.

The numerical methods proposed in [19, 24] address problems of type (**P1**) considering the ℓ_1 regularization to impose piecewise linear solutions [19, 22]:

$$\min_{\boldsymbol{\tau} \in \mathbb{R}^n} \|\mathbf{z} - \Psi(\boldsymbol{\tau})\|_2^2 + \gamma \|\mathrm{L}\boldsymbol{\tau}\|_1, \tag{20}$$

for some $\gamma > 0$. In this section, we adopt the primal-dual formulation with the PDPS algorithm [5, Algorithm 1]. The reason for this choice is that this numerical method is treated specially in the case of non-linear inverse problems such as (**P1**), including the choice of parameters involved in

iterations. In this context, consider the convex conjugate formula applied to the fidelity term:

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

Replacing the variable $\boldsymbol{u}$ by $\Psi(\boldsymbol{\tau})$, we obtain an equivalent minmax formulation of problem (20):

$$\min_{\boldsymbol{\tau}\in\mathbb{R}^n} \max_{\boldsymbol{y}\in\mathbb{R}^n} \gamma\|\mathrm{L}\boldsymbol{\tau}\|_1 + \langle \Psi(\boldsymbol{\tau}), \boldsymbol{y}\rangle - G^*(\boldsymbol{y}). \tag{21}$$

The PDPS algorithm proposes to solve (21) using the same principle of proximal point methods. This algorithm iterates over $k \in \mathbb{N}$:

$$\begin{cases} \boldsymbol{\tau}^{k+1} = \mathrm{prox}_{\sigma_1\gamma\|\cdot\|_1}(\boldsymbol{\tau}^k - \sigma_1\Psi'(\boldsymbol{\tau}^k)y^k) \\ \boldsymbol{y}^{k+1} = \mathrm{prox}_{\sigma_2(G^*-2\langle\Psi(\boldsymbol{\tau}^k),\cdot\rangle)}(\boldsymbol{y}^k - \sigma_2\Psi(\boldsymbol{\tau}^k)), \end{cases} \tag{22}$$

for $\sigma_1, \sigma_2 > 0$. In order to ensure the weak convergence of this algorithm [19, Theorem 1], σ_1 and σ_2 might respect the following inequality established in [19, Example 6]:

$$\sigma_1 \leqslant \frac{1}{\sigma_2 L_\Psi^2 + L_{\Psi'}\rho_{\boldsymbol{y}}/2}$$

where $L_\Psi, L_{\Psi'}$ represent the Lipschitz contants of Ψ and Ψ' respectively. And $\rho_{\boldsymbol{y}}$ the radius of the ball where the iterative sequence (22) converges. We can easily estimate $L_\Psi, L_{\Psi'} \leqslant 1$ by computing the derivative of order 1 and 2 of Ψ. In the following example we compute $\rho_{\boldsymbol{y}} = 1$ by estimating the norm of variable $\boldsymbol{y}$ empirically. We also set $\gamma = 1$ by studying the parameter for which the provided solution $\boldsymbol{\tau}$ is piecewise linear.

Experiment with noiseless signal and local minima: Consider a noiseless read: $\mathbf{z} = \Psi(\bar{\boldsymbol{\tau}})$, where $\mathbf{z}$ and $\bar{\boldsymbol{\tau}}$ are illustrated in Figure 7. Problem (20) is non-convex, implying that a local minimum can not be generalized as a global minimum. For each initial point $(\boldsymbol{\tau}_{\text{initial}}, \boldsymbol{y}_{\text{initial}})$, scheme (22) find a local solution of problem (20). In Figure 8, we analyze the dependence of PDPS solution with respect to these initial points. Let $\boldsymbol{\tau}^i_{\text{Initial}}$, for $i \in \{1, 2, 3, 4\}$ be possible initial points described as follows: $\boldsymbol{\tau}^1_{\text{initial}}$ and $\boldsymbol{\tau}^2_{\text{initial}}$ are constant vectors with values 0.2 and 5 respectively. $\boldsymbol{\tau}^3_{\text{initial}}$ is generated by an uniform distribution between 0 and 5, and $\boldsymbol{\tau}^4_{\text{initial}}$ is a perturbation of the ground truth $\bar{\boldsymbol{\tau}}$. In Table 1 we compare the objective value of different local minima and the correspondent execution time. We observe that only $\boldsymbol{\tau}^4_{\text{initial}}$ is capable to provide the optimal global solution provided by DNA-inverse. In addition, the runtime time of DNA-Inverse is considerably lower then PDPS. In Figure 7, we observe that the DNA-Inverse method closely approximates the original timing profile $\bar{\boldsymbol{\tau}}$. Note that an initial point is not necessary for Algorithm 1.

Experiment with different initiation points: When applying the PDPS method, we find local solutions. These solutions are not arbitrary, they depend on the inverses $\Psi_{\boldsymbol{d}}^{-1}$ defined in Section 5. More specifically, for an initial point $\boldsymbol{\tau}_{\text{initial}}$, consider d_{initial} defined by:

$$\boldsymbol{d}_{\text{initial}} = \operatorname*{arg\,min}_{\boldsymbol{d}\in\{0,1\}^n} \|\boldsymbol{\tau}_{\text{initial}} - \Psi_{\boldsymbol{d}}^{-1}(\mathbf{z})\|_2^2.$$

Results exposed in Figure 8 indicate that the PDPS algorithm results in a piecewise vector that approximates $\Psi^{-1}_{\boldsymbol{d}_{\text{initial}}}(\mathbf{z})$. This fact suggests that we can adapt the initialization of DNA-Inverse, Algorithm 1, for the PDPS. The adapted algorithm loops as follows: Step 1 - For each $d \in \widetilde{\mathcal{D}}$, we set

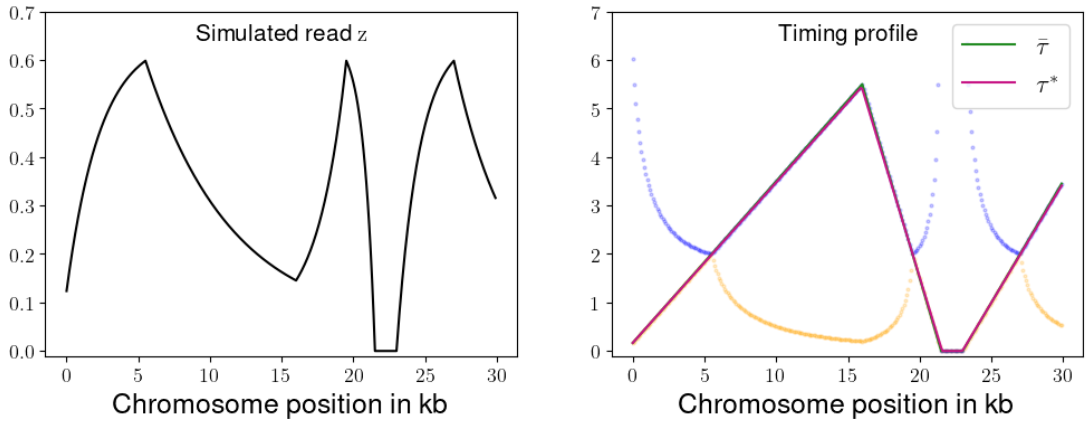

Figure 7: (left) Simulated read $z = \Psi(\bar{\tau})$. (right) Solution $\boldsymbol{\tau}^*$ obtained via the DNA-Inverse method (in pink) compared to ground truth $\bar{\boldsymbol{\tau}}$ (in green). In orange, $\Psi_{\mathbf{0}}^{-1}(\mathbf{z})$, and in blue, $\Psi_{\mathcal{I}}^{\mathbf{1}}(\mathbf{z})$. We observe a close resemblance between the piecewise linear vectors $\bar{\boldsymbol{\tau}}$ and $\boldsymbol{\tau}^*$.

Method	$\boldsymbol{\tau}_{\text{initial}}$	Objective value of local minima	Execution time
PDPS	$\boldsymbol{\tau}^1_{\text{initial}}$	0.18	45s
PDPS	$\boldsymbol{\tau}^2_{\text{initial}}$	0.58	45s
PDPS	$\boldsymbol{\tau}^3_{\text{initial}}$	0.12	46s
PDPS	$\boldsymbol{\tau}^4_{\text{initial}}$	0.073	40s
DNA-Inverse	$--$	0.073	$7s$

Table 1: Comparison between objective value and execution time for different initiation points in the case of a noiseless signal z illustrated in Figure 7. The convergence criterion for PDPS limits distance between two consecutive iterates: $\|\tau_k - \tau_{k+1}\| \leqslant 1\text{e}{-5}$.

as initial point the vector $\Psi_d^{-1}(\mathbf{z})$; Step 2 - Compare the objective values for each $d \in \widetilde{\mathcal{D}}$ and chose the smaller one. We call this strategy *Adapted PDPS* and its output: $(\tau^*_{\text{PDPS}}, d^*_{\text{PDPS}})$. In Figure 9 we observe the similarity between solutions for three different elements of the real data-set of yeast reads. For all these cases, the optimal $d^*_{\text{PDPS}} = d^*_{\text{DNA-Inverse}}$ and $\tau^*_{\text{PDPS}} \approx \tau^*_{\text{DNA-Inverse}}$.

As illustrated in Figure 9, the *Adapted PDPS* has shown to be efficient in providing a global solution to problem (20). Nevertheless, the main drawback of this adaptation is the execution time. The convergence of the proximal algorithm is slow, resulting in a significantly high total execution time, as displayed in the Figure 10 and in Table 2.

Method	Mean execution time	Median execution time
DNA-Inverse	42s	10s
PDPS	272s	163s

Table 2: Comparison of execution time between DNA-Inverse and Adapted PDPS

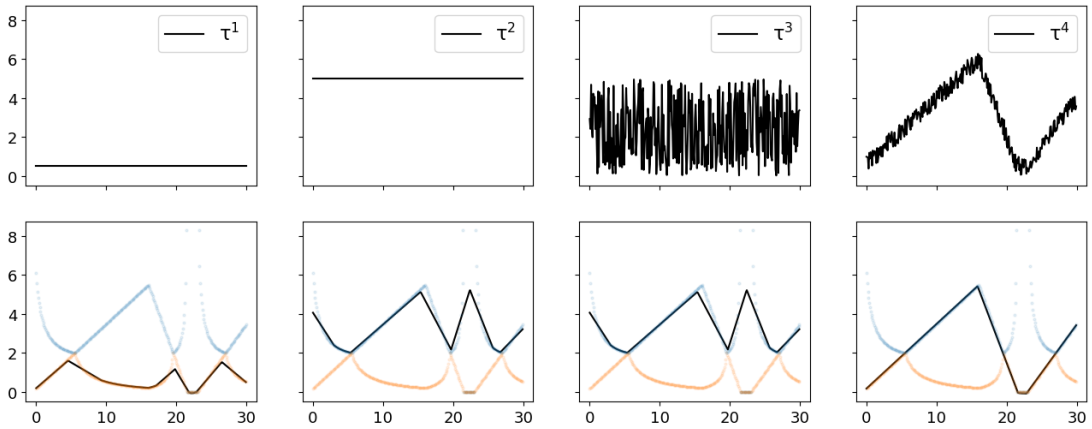

Figure 8: Illustration of the results of PDPS algorithm for different initial points. (above): different values of $\boldsymbol{\tau}^i_{\text{initial}}$ for $i \in \{1, 2, 3, 4\}$. (below): solutions of the PDPS algorithm with the correspondent initial points. In orange we observe $\mathbf{z^0} = \Psi_{\mathbf{0}}^{-1}(\mathbf{z})$ and in blue $\mathbf{z^1} = \Psi_{\mathbf{1}}^{-1}(\mathbf{z})$.

6.2 Advantages of DNA-Inverse with respect to other methods:

The primary advantage of the DNA-Inverse algorithm over previous works [14, 15, 16] is its ability to detect any replication event. Unlike previous methods, which only detected replication origins that had been activated before the beginning of the experiment, the DNA-Inverse algorithm is capable of detecting all replication events. In this context, various cases of interest, which are prevalent in the database, were previously excluded from the statistics, as illustrated in Figure 11.

The effect of ℓ_1 regularization in (10) is to introduce a bias that tends to yield lower speed values because it affects the angular coefficient of lines [26]. To mitigate this drawback, the DNA-Inverse method enables an enhancement of estimation at low cost. Note that after correctly detecting the variable $\boldsymbol{d}^*$, the problem of finding the time profile $\boldsymbol{\tau}^*$ consists in fitting a piecewise linear function in a noisy data, a task for which various methods can be applied [27, 28]. This possibility is promising for achieving velocity estimation with unprecedented accuracy in future studies.

7 Conclusion

In this work, we analyzed DNA replication in single-molecule from the perspective of a nonlinear and non-convex inverse problem. We developed the model called DNA-Inverse, which effectively integrates different possibilities of local solutions, eliminating the necessity to solve the problem with multiple initial points. We studied theoretical results demonstrating the coherence of the proposed inverse problem and compare its results with state-of-the-art methods, highlighting its advantage in terms of execution time. Additionally, we showed how the results from this formulation outperform previous works, expanding the amount of data and replication events that can be analyzed by biologists.

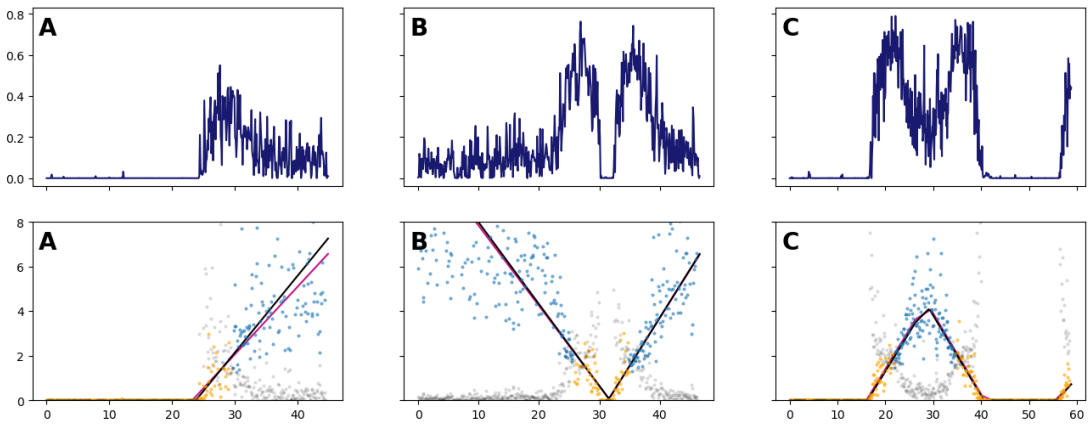

Figure 9: (Top) Different reads in blue representing the following events: A. Progression of a fork to the right; B. Initiation; C. Termination. (Bottom) Solution $\boldsymbol{\tau}^*_{\text{DNA-Inverse}}$ (in pink) and $\boldsymbol{\tau}^*_{\text{PDPS}}$ (in black) for the different reads. Colored points are selected by the the solution d^* that is the same for both methods. For $i \in \{1, ..., n\}$, points in blue are the inverse $\psi_0^{-1}(z_i)$ where $d_i^* = 0$, and points in orange are the inverse $\psi_1^{-1}(z_i)$ when $d_i^* = 1$. Black points are those not selected for the solution d^*. We emphasize the similarity of the time profiles for PDPS and DNA-Inverse, and the fact that the solutions in variable $\boldsymbol{d}$ are the same.

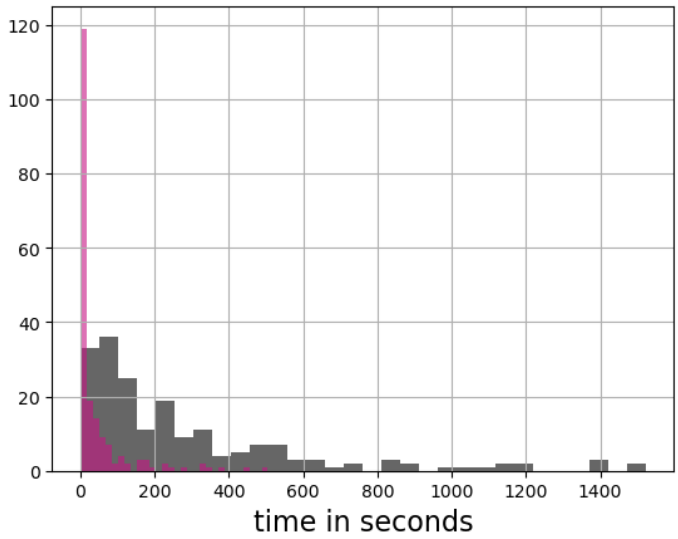

Figure 10: Comparison between execution time distribution for DNA-Inverse (pink) and Adapted PDPS (black). The median execution time of DNA-Inverse is 16 times smaller than that of PDPS.

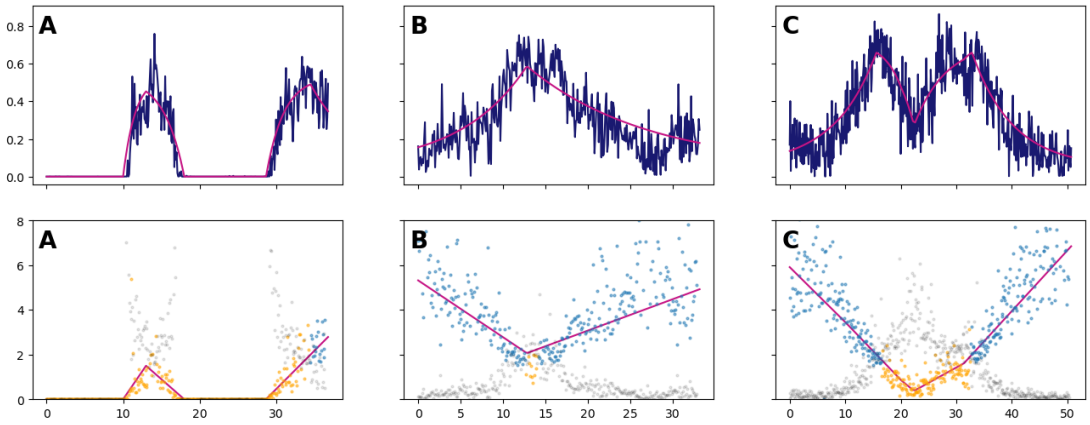

Figure 11: (Top) Different reads in blue representing the following events: A. (left) Termination (right) fork progresses to the right; B. Initiation; C. Initiation. The line in pink represent the approximation $\Psi_{d*}(\tau^*)$ given by the optimal solution of the DNA-Inverse method. (Bottom) Solution $\boldsymbol{\tau}^*$ for DNA-Inverse (in pink). Colored points are selected by the the solution d^*. For $i \in \{1, ..., n\}$, points in blue are the inverse $\psi_0^{-1}(z_i)$ where $d_i^* = 0$, and points in orange are the inverse $\psi_1^{-1}(z_i)$ when $d_i^* = 1$. Black points are inverses not selected by the optimal d^*. All these events could not be detected in previous works.

References

[1] H. Gaillard, T. García-Muse, and A. Aguilera. "Replication stress and cancer". In: *Nature Reviews Cancer* 15.5 (2015), pp. 276–289. DOI: 10.1038/nrc3916. URL: https://doi.org/10.1038/nrc3916.

[2] J. Jauhiainen, A. Seppänen, and T. Valkonen. "Mumford–Shah regularization in electrical impedance tomography with complete electrode model". In: *Inverse Problems* 38.6 (2022), p. 065004. DOI: 10.1088/1361-6420/ac5f3a. URL: https://dx.doi.org/10.1088/1361-6420/ac5f3a.

[3] Y. Shechtman et al. "Sparsity-based super-resolution and phase-retrieval in waveguide arrays". In: *Opt. Express* 21.20 (2013), pp. 24015–24024. URL: http://www.opticsexpress.org/abstract.cfm?URI=oe-21-20-24015.

[4] Y. Shechtman, A. Beck, and Y. C. Eldar. "GESPAR: Efficient Phase Retrieval of Sparse Signals". In: *IEEE Transactions on Signal Processing* 62.4 (2014), pp. 928–938.

[5] C. Clason, S. Mazurenko, and T. Valkonen. "Acceleration and Global Convergence of a First-Order Primal-Dual Method for Nonconvex Problems". In: *SIAM Journal on Optimization* 29.1 (2019), pp. 933–963. DOI: 10.1137/18M1170194. URL: https://doi.org/10.1137/18M1170194.

[6] H. Li et al. "NETT: solving inverse problems with deep neural networks". In: *Inverse Problems* 36.6 (2020), p. 065005. DOI: 10.1088/1361-6420/ab6d57. URL: https://dx.doi.org/10.1088/1361-6420/ab6d57.

[7] A. Chambolle and T. Pock. "A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging". In: *Journal of Mathematical Imaging and Vision* 40.1 (2011), pp. 120–145.

[8] H. Kunze, D. L. Torre, and M. R. Galán. "Optimization methods in inverse problems and applications to science and engineering". In: *Optimization and Engineering* 22.4 (2021), pp. 2151–2158. DOI: 10.1007/s11081-021-09688-y. URL: https://doi.org/10.1007/s11081-021-09688-y.

[9] Y. Wang, W. Yin, and J. Zeng. "Global Convergence of ADMM in Nonconvex Nonsmooth Optimization". In: *Journal of Scientific Computing* 78.1 (2019), pp. 29–63. DOI: 10.1007/s10915-018-0757-z. URL: https://doi.org/10.1007/s10915-018-0757-z.

[10] Y. Wang, J. Lacotte, and M. Pilanci. "The Hidden Convex Optimization Landscape of Regularized Two-Layer ReLU Networks: an Exact Characterization of Optimal Solutions". In: *International Conference on Learning Representations*. 2020. URL: https://api.semanticscholar.org/CorpusID:251647158.

[11] T. Ergen and M. Pilanci. *The Convex Landscape of Neural Networks: Characterizing Global Optima and Stationary Points via Lasso Models*. 2023.

[12] D. Bertsimas, A. King, and R. Mazumder. "Best Subset Selection via a Modern Optimization Lens". In: *The Annals of Statistics* 44.2 (2016), pp. 813–852. DOI: 10.1214/15-AOS1388.

[13] M. Hennion et al. "FORK-seq: replication landscape of the Saccharomyces cerevisiae genome by nanopore sequencing". In: *Genome Biology* 21.1 (2020), p. 125. DOI: 10.1186/s13059-020-02013-3. URL: https://doi.org/10.1186/s13059-020-02013-3.

[14] B. T. et al. "Genome-wide mapping of individual replication fork velocities using nanopore sequencing". In: *Nature Communications* 13 (2022). DOI: 10.1038/s41467-022-31012-0.

[15] C. Lage et al. “Codage espace-échelle parcimonieux en présence de bruit non-gaussien. Application à l’analyse de la réplication de l’ADN en molécule unique”. In: *GRETSI* (2023).

[16] C. Lage et al. *Space-Scale Hybrid Continuous-Discrete Sliding Frank-Wolfe Method*. 2023.

[17] S. Mallat and Z. Zhang. “Matching pursuits with time-frequency dictionaries”. In: *IEEE Transactions on Signal Processing* 41.12 (1993), pp. 3397–3415. DOI: 10.1109/78.258082.

[18] Y. Shechtman and et al. “Phase retrieval with application to optical imaging: a contemporary overview”. In: *IEEE Signal Processing Magazine* 32.3 (2015), pp. 87–109.

[19] T. Valkonen. “First-Order Primal–Dual Methods for Nonsmooth Non-convex Optimisation”. In: (2021). Ed. by K. Chen et al., pp. 1–42. DOI: 10.1007/978-3-030-03009-4_93-1. URL: https://doi.org/10.1007/978-3-030-03009-4_93-1.

[20] P. Boufounos. “Greedy sparse signal reconstruction from sign measurements”. In: 2009, pp. 1305–1309. DOI: 10.1109/ACSSC.2009.5469926.

[21] L. Rencker et al. “Sparse Recovery and Dictionary Learning From Nonlinear Compressive Measurements”. In: *IEEE Transactions on Signal Processing* 67.21 (2019), pp. 5659–5670. DOI: 10.1109/TSP.2019.2941070.

[22] A. Ali and R. J. Tibshirani. *The Generalized Lasso Problem and Uniqueness*. 2019.

[23] J. T. Ryan J. Tibshirani. “The solution path of the generalized lasso”. In: *The Annals of Statistics* 39 (), pp. 1335, 1371.

[24] G. Li and T. K. Pong. “Global Convergence of Splitting Methods for Nonconvex Composite Optimization”. In: *SIAM Journal on Optimization* 25.4 (2015), pp. 2434–2460. DOI: 10.1137/140998135. URL: https://doi.org/10.1137/140998135.

[25] F. Bian, J. Liang, and X. Zhang. “A stochastic alternating direction method of multipliers for non-smooth and non-convex optimization”. In: *Inverse Problems* 37.7 (2021), p. 075009. DOI: 10.1088/1361-6420/ac0966. URL: https://dx.doi.org/10.1088/1361-6420/ac0966.

[26] D. Vidaurre, C. Bielza, and P. Larrañaga. “A Survey of L1 Regression”. In: *International Statistical Review* 81.3 (2013), pp. 361–387. DOI: https://doi.org/10.1111/insr.12023. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/insr.12023.

[27] A. Magnani and S. P. Boyd. “Convex piecewise-linear fitting”. In: *Optimization and Engineering* 10.1 (2009), pp. 1–17. DOI: 10.1007/s11081-008-9045-3. URL: https://doi.org/10.1007/s11081-008-9045-3.

[28] V. E. McZgee and W. T. Carleton. “Piecewise Regression”. In: *Journal of the American Statistical Association* 65.331 (1970), pp. 1109–1124. DOI: 10.1080/01621459.1970.10481147. URL: https://doi.org/10.1080/01621459.1970.10481147.