# Measurement Errors Make the Partial Digest Problem NP-hard

Mark Cieliebak[1] and Stephan Eidenbenz[2]

[1] Institute of Theoretical Computer Science, ETH Zurich, cieliebak@inf.ethz.ch
[2] Los Alamos National Laboratory*, eidenben@lanl.gov

**Abstract.** The Partial Digest problem asks for the coordinates of $m$ points on a line such that the pairwise distances of the points form a given multiset of $\binom{m}{2}$ distances. Partial Digest is a well-studied problem with important applications in physical mapping of DNA molecules. Its computational complexity status is open. Input data for Partial Digest from real-life experiments are always prone to error, which suggests to study variations of Partial Digest that take this fact into account. In this paper, we study the computational complexity of the variation of Partial Digest in which each distance is known only up to some error, due to experimental inaccuracies. The error can be specified either by some additive offset or by a multiplicative factor. We show that both types of error make the Partial Digest problem strongly NP-complete, by giving reductions from 3-Partition. In the case of relative errors, we show that the problem is hard to solve even for constant relative error.

## 1 Introduction

The Partial Digest problem is perhaps *the* classic combinatorial problem from computational biology with applications in DNA sequencing. Despite considerable research efforts in the past twenty years, its computational complexity is still an open problem. In the Partial Digest problem we are given a multiset $D$ of distances and are asked to find coordinates of points on a line, i.e., a point set $P$, such that $D$ is exactly the multiset[3] of all pairwise distances of these points. In this case, we say that $D$ is the distance multiset of point set $P$. A formal definition of the problem is as follows.

**Definition 1** (Partial Digest)**.** *Given an integer $m$ and a multiset of $k = \binom{m}{2}$ positive integers $D = \{d_1, \ldots, d_k\}$, is there a set of $m$ integers $P = \{p_1, \ldots, p_m\}$ such that $\{|p_i - p_j| \mid 1 \leq i < j \leq m\} = D$?*

For example, if $D = \{2, 5, 7, 7, 9, 9, 14, 14, 16, 23\}$, then $P = \{0, 7, 9, 14, 23\}$ is one feasible solution (cf. Figure 1).

* Work partially done while M. Cieliebak was visiting LANL. LA-UR-03:6621.

[3] We will denote multisets like sets, since the fact of being a multiset is not crucial for our purposes.

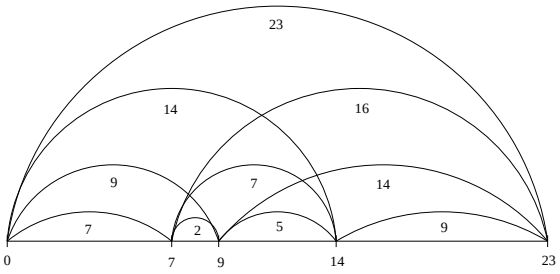


**Fig. 1.** Example for Partial Digest

**Previous Work**

Intriguingly, the computational complexity of this seemingly straight-forward combinatorial puzzle is a long–standing open problem, and it appears in its pure combinatorial formulation already in the 1930's in the area of X–ray crystallography (acc. to [16]). The problem is also known as "turnpike problem", where we are given the pairwise distances of cities along a highway, and we want to find their ordering along the road [4]. The Partial Digest problem can be solved in pseudo–polynomial time [10, 13], and there exists a backtracking algorithm (for exact or erroneous data) that has expected running time polynomial in the number of distances [16, 17], but exponential worst case running time [20]. The Partial Digest problem can be formalized by cut grammars, which have one additional symbol $\delta$, the *cut*, that is neither a non–terminal nor a terminal symbol [14], and the problem is closely related to the theory of homometric sets[4] [16]. Finally, if the points in a solution do not have to be on a line, but only in $d$–dimensional space, then the problem is NP-hard for some $d \geq 2$ [16]. However, for the original Partial Digest problem, neither a polynomial–time algorithm nor a proof of NP-hardness is known [2, 4, 11, 12, 15].

**Biological Background**

Partial Digest has several applications; the classical and most prominent is in the study of the structure of DNA molecules. More precisely, given a large DNA molecule (sequence of nucleotides `A`, `C`, `G`, and `T`), restriction enzymes can be used to generate a physical map of the molecule. A restriction enzyme cuts a DNA molecule at specific patterns, the restriction sites. For instance, the enzyme Eco RI cuts occurrences of the pattern `GAATTC` into `G` and `AATTC`. Under appropriate experimental conditions, *all* fragments between each two restriction sites are created. This process is called *partial digestion*. The lengths of the fragments (i.e., their number of nucleotides) are then measured by using gel electrophoresis, a standard technique in molecular biology. This leaves us with the multiset of distances between all restriction sites, and the objective is to

[4] Two (non–congruent) sets of points are homometric if they generate the same multiset of pairwise distances.

reconstruct the original ordering of the fragments in the DNA molecule, which is the Partial Digest problem.

**Erroneous Data**

In real-life, partial digestion experiments cannot be conducted under ideal conditions as outlined above, and thus errors occur in the data. In fact, there is no such thing as error–free data, and typically four types of errors occur [5, 6, 8, 17]: *additional fragments*, for instance through contamination of the probe with unrelated biological material; *missing fragments*, due to partial cleavage errors, or because of small fragments that remain undetected by gel electrophoresis; *incorrect fragment lengths*, due to the fact that fragment lengths cannot be determined exactly using gel electrophoresis; and, finally, *wrong multiplicities*, due to the intrinsic difficulty to determine the proper multiplicity of a distance by gel electrophoresis[5].

Algorithms for Partial Digest with inaccurate data have been studied intensively in the literature [5, 8, 17], and different error models have been designed, e.g. for measurement errors that are logarithmic in the size of the fragment length [18, 19] or for intervals of absolute errors [1, 17]. Optimization variations of Partial Digest where fragments are either omitted or added in the data, and the number of errors has to be minimized, are known to be NP-hard or hard to approximate, respectively [3].

In this work we will focus on the third type of error, where the lengths of fragments can be erroneous (*measurement errors*). In partial digestion experiments all measurements of fragment lengths are prone to inaccuracies: Using gel electrophoresis, measurement errors within a range of up to 5 percent of the fragment length can occur [5, 6, 17].

Many experimental variations of partial digest experiments have been studied, see [9] for a survey; and for more detailed discussions on the problem, see [12] and [15].

**Definitions and Results**

In this paper, we study the computational complexity of Partial Digest in the presence of measurement errors, where we allow both additive or multiplicative errors.

We start with additive errors. The Partial Digest problem is known to be strongly NP-hard if additive error bounds that can be even zero can be assigned to each distance *individually* [9, 16]. However, this does not model reality appropriately, since in real–life data we cannot assume that even one single fragment length can be measured exactly. Therefore, we study the computational complexity of the variation of Partial Digest where *all* measurements are prone to some non–zero error. Moreover, we refrain from individual error bounds, and study the variation where all measurements are prone to *the same additive non–zero* error $\delta$. More precisely, we say that value $v$ matches a distance $d$ up to additive error $\delta$ if $|v - d| \leq \delta$; moreover, a multiset $D$ is a distance multiset of a point set $P$ up to additive error $\delta$, if there is a bijective function

[5] The multiplicity of a fragment is determined from the intensity of the corresponding band in the gel.

$f : D \to \Delta(P)$ such that each distance $d \in D$ matches value $f(d)$ up to error $\delta$; here, $\Delta(P) = \{|p_j - p_i| \mid 1 \leq i < j \leq n\}$ denotes the multiset of pairwise distances in $P$. The PD–AbsError problem is defined as follows.

**Definition 2** (PD–AbsError)**.** *Given an integer $m$, a multiset $D$ of $k = \binom{m}{2}$ positive integers, and an integer error bound $\delta > 0$, is there a set $P$ of $m$ points on a line such that $D$ is the distance multiset of $P$ up to additive error $\delta$?*

We show in Section 2 that PD–AbsError is strongly NP-complete, by giving a reduction from 3-Partition.

We then turn to the case of multiplicative errors. We say that distance $d$ matches a value $x$ up to *multiplicative error* $\varepsilon > 0$ if $d(1-\varepsilon) \leq x \leq d(1+\varepsilon)$. Observe that this definition is not symmetric, i.e., if $d$ matches $x$ up to error $\varepsilon$, then this does *not* in general imply that $x$ matches $d$ (in contrast to the definition of additive errors, which is symmetric). A multiset $D$ is a distance multiset of point set $P$ up to multiplicative error $\varepsilon$ if there is a bijective function $f : D \to \Delta(P)$ such that each distance $d \in D$ matches value $f(d)$ up to multiplicative error $\varepsilon$. The PD–RelError problem is defined as follows.

**Definition 3** (PD–RelError)**.** *Given an integer $m$, a multiset $D$ of $k = \binom{m}{2}$ positive integers, and a rational error $\varepsilon > 0$, is there a set $P$ of $m$ points on a line such that $D$ is the distance multiset of $P$ up to multiplicative error $\varepsilon$?*

We show in Section 3 that PD–RelError is strongly NP-complete, even for constant error, by using a similar reduction as for PD–AbsError.

## 2 Strong NP-completeness of PD–AbsError

In this section, we show that PD–AbsError is strongly NP-complete, by giving a reduction from 3-Partition, which is the following problem: Given $3n$ positive integers $q_1, \ldots, q_{3n}$ and an integer $h$ such that $\sum_{i=1}^{3n} q_i = nh$ and $\frac{h}{4} < q_i < \frac{h}{2}$ for $i \in \{1, \ldots, 3n\}$, are there $n$ disjoint triples of $q_i$'s such that each triple adds up to $h$? The 3-Partition problem is NP-complete in the strong sense [7]. Observe that $\frac{h}{4} < q_i < \frac{h}{2}$ already implies that each subset of the $q_i$'s that adds up to $h$ must have exactly three elements.

The idea of the reduction is as follows. Given an instance $q_1, \ldots, q_{3n}$ and $h$ of 3-Partition, we define a multiset of distances $D$ and an additive error $\delta = \frac{h}{4}$ that form an instance of PD–AbsError. Our construction is based on the following observation: If there is a solution for the 3-Partition instance, then we can arrange the $q_i$'s such that triples of adjacent $q_i$'s sum up to $h$. If we sum up, say, 25 adjacent $q_i$, then we sum over at least 7 complete triples that each have sum $h$, plus some few (up to four) additional $q_i$'s at the beginning and the end. In the special and trivial case that all $q_i$'s have exactly value $\frac{h}{3}$, we can easily determine the exact sum of the 25 values. However, in a given instance of 3-Partition typically not all $q_i$'s will have value $\frac{h}{3}$. However, they have "approximately" value $\frac{h}{3}$, since they satisfy $\frac{h}{4} < q_i < \frac{h}{2}$ by definition. In the proof of the following theorem, we will use additive error $\delta$ to "close the gap" between $\frac{h}{3}$ and the true values of the $q_i$'s.

**Theorem 4.** PD–AbsError *is strongly* NP*-complete.*

*Proof.* The problem PD–AbsError is in NP: Given a candidate point set $P$, we sort all distances between any two points in $P$, and all distances in $D$; then $P$ is a solution if error $\delta$ is sufficient to match the $i$-th distance from $P$ to the $i$-th distance from $D$.

To prove strong NP-hardness, we give a reduction from 3-Partition. Given an instance of 3-Partition, i.e., integers $q_1, \ldots, q_{3n}$ and integer $h$, we define a distance multiset $D$ and an additive error $\delta$ that are an instance of PD–AbsError. There will be a solution for this instance if and only if there is a solution for the 3-Partition instance. Parallel to the definition of $D$, we show already the "if" direction of the previous statement: To this end, we assume that the 3-Partition can be solved, i.e., there are $n$ triples $T_1, \ldots, T_n$ of $q_i$'s that each sum up to $h$. We show how to construct a point set $P$ that is a solution for the PD–AbsError instance, i.e., $P$ matches $D$ up to additive error $\delta$. The opposite direction ("only if") is shown in a second step. We want to stress at this point that although the definition of $D$ and the construction of $P$ are presented simultaneously, the definition of $D$ itself does *not* rely on the fact that there exists a solution for the 3-Partition instance.

We assume that $\frac{h}{12}$ is integer. Otherwise, we can achieve this by simply multiplying all values $q_i$ and $h$ by 12. Moreover, we assume w.l.o.g. that the values $q_1, \ldots, q_{3n}$ are ordered such that the three $q_i$'s that belong to the same triple $T_j$ in a solution are adjacent, i.e., $T_1 = (q_1, q_2, q_3), T_2 = (q_4, q_5, q_6)$, and so on. Finally, we assume that the elements in each $T_i$ are sorted in ascending order, i.e., $q_1 \leq q_2 \leq q_3, q_4 \leq q_5 \leq q_6$, and so on. This ordering allows us to derive a set of inequalities for the $q_i's$. Let $(q_{3k+1}, q_{3k+2}, q_{3k+3})$ be a triple that sums up to $h$, for $0 \leq k \leq n-1$. Then $q_{3k+1} \leq \frac{h}{3}$, since $q_{3k+1}$ is the smallest of the three elements in the triple, and not all of them can be greater than $\frac{h}{3}$. Similarly, $\frac{h}{3} \leq q_{3k+3}$. With $q_{3k+1} + q_{3k+2} = h - q_{3k+3}$, we have $q_{3k+1} + q_{3k+2} \leq h - \frac{h}{3} = \frac{2h}{3}$. In combination with the restriction $\frac{h}{4} < q_i < \frac{h}{2}$ (from the definition of 3-Partition) and $H := \frac{h}{12}$, this yields the following inequalities:

$$\begin{aligned}
3H &< q_{3k+1} && \leq 4H \\
3H &< q_{3k+2} && < 6H \\
4H &\leq q_{3k+3} && < 6H \\
6H &< q_{3k+1} + q_{3k+2} && \leq 8H \\
8H &\leq q_{3k+2} + q_{3k+3} && < 12H \\
12H &= q_{3k+1} + q_{3k+2} + q_{3k+3}
\end{aligned} \tag{1}$$

We will use these inequalities later to derive upper and lower bounds for the additive error that we need to apply to our distances in order to guarantee the existence of a solution for the PD–AbsError instance.

Before we define our distances, we need to introduce the *level* of a distance: For a point set $P$, we say that a distance $d$ between two points has *level* $\ell$ if it spans $\ell - 1$ further points, and we say that distance $d$ is an *atom* if it has level 1. E.g. in Figure 1, distance 5 is an atom, and distance 16 has level 3.

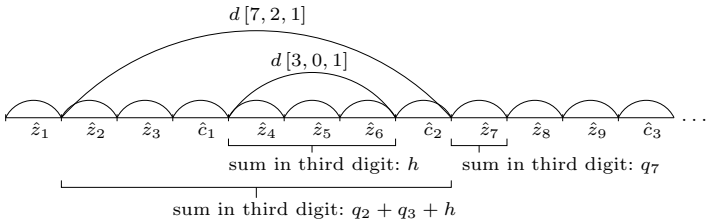


**Fig. 2.** Atoms and distances in multiset $D$.

In the following, we will use a vector representation for large numbers that will allow to add up the numbers digit by digit. The numbers are expressed in the number system of some base $Z$. We denote by $\langle a_1, \ldots, a_n \rangle$ the number $\sum_{i=1}^{n} a_i Z^{n-i}$; we say that $a_i$ is the $i$-th digit of this number. In our proofs, we will choose base $Z$ large enough such that the additions that we will perform do not lead to carry–overs from one digit to the next. Hence, we can add numbers digit by digit. The same holds for scalar multiplications. For example, having base $Z = 29$ and numbers $\alpha = \langle 3, 5, 1 \rangle$ and $\beta = \langle 2, 1, 0 \rangle$, then $\alpha + \beta = \langle 5, 6, 1 \rangle$ and $3 \cdot \alpha = \langle 9, 15, 3 \rangle$.

We now define our instance of PD–AbsError and show at the same time how to construct a solution for this instance. Let $c = n^2 \cdot h^2$. Moreover, define error $\delta := 3H$. The distances are expressed as numbers with base $Z = 10nc$, and each distance consists of three digits. The first digit will denote the *level* of a distance (the meaning of the other two digits will become clear soon).

First we define $4n-1$ distances that will turn out to be atoms in our solution: $z_i = \langle 1, 0, q_i \rangle - \delta$, for $1 \le i \le 3n$, and $c_i = \langle 1, c, 0 \rangle - \delta$, for $1 \le i \le n-1$. Observe that operation "$-\delta$" only affects the last digit (and in fact, we could have defined $z_i$ by $\langle 1, 0, q_i - \delta \rangle$ instead), since we choose base $Z$ sufficiently large.

Using these distances, we can already define a "solution" $P$ for distance multiset $D$ (although we are not yet finished defining $D$; in fact, we will construct $D$ in the following such that it matches point set $P$ up to additive error $\delta$): Let $\hat{z}_i = z_i + \delta$ for $1 \le i \le 3n$, and $\hat{c}_i = c_i + \delta$ for $1 \le i \le n-1$. Observe that each $\hat{z}_i$ has exactly value $q_i$ in its third digit. We call these values *z–pseudoatoms* or *c–pseudoatoms*, respectively, and use them to define a point set $P = \{p_1, \ldots, p_{4n}\}$ by specifying the pairwise distances between the points: Starting in 0, the points have distances $\hat{z}_1, \hat{z}_2, \hat{z}_3, \hat{c}_1, \hat{z}_4, \hat{z}_5, \hat{z}_6, \hat{c}_2, \ldots, \hat{c}_{n-1}, \hat{z}_{3n-2}, \hat{z}_{3n-1}, \hat{z}_{3n}$, i.e., we alternate blocks of three $z$–pseudoatoms and one $c$–pseudoatom, starting and ending with a block of three $z$–pseudoatoms (see Figure 2).

We now show level by level how the distances in $D$ are defined, and that additive error $\delta$ (which is $3H$) is sufficient to make all distances from $D$ match some distance between points in $P$.

By construction of $P$, the distances of level 1 are the pseudoatoms, and they match the corresponding $z_i$'s and $c_i$'s up to additive error $\delta$.

To denote the distances of higher levels we use notation $d\,[\ell, j, k]$, for appropriate parameters $\ell, j$ and $k$. These names already indicate the values of the

three digits of a distance: Distance $d[\ell, j, k]$ will have value $\ell$ in the first digit, which will be the level of the distance in our point set $P$. The second digit of the distance has value $j \cdot c$, which denotes that this distance will be used to span $j$ $c$–pseudoatoms (and $\ell - j$ $z$–pseudoatoms) in our point set $P$. For instance, in Figure 2 distance $d[7, 2, 1]$ spans the two pseudoatoms $\hat{c}_1$ and $\hat{c}_2$ (and five $\hat{z}_i$'s). Finally, the third digit of distance $d[\ell, j, k]$ has value $k \cdot h$ plus some "small offset", which will be a multiple of $H$. Here, $k$ specifies how many *complete* blocks of three adjacent $z$–pseudoatoms the distance spans in $P$ (recall that such a block corresponds to three $q_i$'s that sum up to exactly $h$). In the following, we show how to choose these offsets in the third digit such that our point set $P$ matches distance multiset $D$ up to additive error $\delta$.

First consider distances of level 2 in $P$, i.e., two points $p_i, p_{i+2} \in P$ with one point $p_{i+1}$ in between. There are four possibilities for the two pseudoatoms between these two points, for some $0 \leq k \leq n-1$: Case 1: $\hat{z}_{3k+1}$ and $\hat{z}_{3k+2}$; Case 2: $\hat{z}_{3k+2}$ and $\hat{z}_{3k+3}$; Case 3: $\hat{z}_{3k+3}$ and $\hat{c}_k$; and Case 4: $\hat{c}_k$ and $\hat{z}_{3k+1}$.

For the first case, the two pseudoatoms sum up to 2 in the first and to 0 in the second digit. For the third digit of the sum, recall that $\hat{z}_{3k+1}$ has value $q_{3k+1}$ in its third digit, and $\hat{z}_{3k+2}$ has value $q_{3k+2}$ in its third digit. Hence, inequalities (1) yield that the third digit of $\hat{z}_{3k+1} + \hat{z}_{3k+2}$ is bounded below by $6H$ and bounded above by $8H$. We define a distance $d[2, 0, 0] := \langle 2, 0, 9H \rangle$. Obviously, we can span the two pseudoatoms by this distance if we apply at most error $\delta$ (recall that $\delta = 3H$). Observe that we could have chosen other values for the third digit of $d[2, 0, 0]$, namely any value between $5H$ and $9H$ (which still allows to match the bounds using additive error $\delta$). Here, we chose value $9H$, since we will use that same distance to cover the two pseudoatoms in Case 2 as well (see below).

Case 1 occurs exactly $n$ times in our point set $P$, once for each block of three $z$–pseudoatoms. Hence, we let distance $d[2, 0, 0]$ be $n$ times in our distance multiset $D$.

Case 2 is similar to Case 1: The third digit of $\hat{z}_{3k+2} + \hat{z}_{3k+3}$ is bounded below by $8H$ and bounded above by $12H$, using again inequalities (1). Like before, this case occurs $n$ times, and we can use $n$ *additional* distances $d[2, 0, 0]$ in $D$ to span such two pseudoatoms up to error $\delta$. Thus, in total we have $2n$ distances $d[2, 0, 0]$ in $D$ that arise from the first two cases.

For the remaining two cases of two pseudoatoms, the last digit of the two pseudoatoms is at least $4H$ and at most $6H$ in Case 3, and at least $3H$ and at most $4H$ in Case 4. Moreover, in both cases the first digit of the sum is 2 and the second digit is $c$, and both cases occur exactly $n - 1$ times. Hence, we can define distance $d[2, 1, 0] := \langle 2, c, 4H \rangle$ and enclose it $2(n-1)$ times in $D$, in order to cover these pairs of pseudoatoms, again up to additive error $\delta$.

Before we specify the distances of higher level, we introduce a graphical representation of pseudoatoms: Each $z$–pseudoatom is represented by a $\bullet$, and each $c$–pseudoatom by a $|$. This allows us to depict sequences of pseudoatoms without referring to their exact names. E.g. pseudoatoms $\hat{z}_3\hat{c}_1\hat{z}_4\hat{z}_5\hat{z}_6\hat{c}_2$ yield $\bullet|\bullet\bullet\bullet|$, and the four cases of two adjacent pseudoatoms above can be represented

| level $\ell$ | pseudoatoms | multiplicity | lower bound | upper bound | distance name | distance value |
|---|---|---|---|---|---|---|
| 2 | ●● | $n$ | $6H$ | $8H$ | $d\,[2,0,0]$ | $\langle 2,0,9H\rangle$ |
| | ●● | $n$ | $8H$ | $12H$ | $d\,[2,0,0]$ | |
| | ●\| | $n-1$ | $4H$ | $6H$ | $d\,[2,1,0]$ | $\langle 2,c,4H\rangle$ |
| | \|● | $n-1$ | $3H$ | $4H$ | $d\,[2,1,0]$ | |
| 3 | ●●● | $n$ | $12H$ | $12H$ | $d\,[3,0,1]$ | $\langle 3,0,12H\rangle+\delta$ |
| | \|●● | $n-1$ | $6H$ | $8H$ | $d\,[3,1,0]$ | $\langle 3,c,9H\rangle$ |
| | ●\|● | $n-1$ | $7H$ | $10H$ | $d\,[3,1,0]$ | |
| | ●●\| | $n-1$ | $8H$ | $12H$ | $d\,[3,1,0]$ | |
| 4 | ●●\|● | $n-1$ | $11H$ | $16H$ | $d\,[4,1,0]$ | $\langle 4,c,13H\rangle$ |
| | ●\|●● | $n-1$ | $10H$ | $14H$ | $d\,[4,1,0]$ | |
| | ●●●\| | $n-1$ | $12H$ | $12H$ | $d\,[4,1,1]$ | $\langle 4,c,12H\rangle$ |
| | \|●●● | $n-1$ | $12H$ | $12H$ | $d\,[4,1,1]$ | |
| 5 | ●●\|●● | $n-1$ | $14H$ | $20H$ | $d\,[5,1,0]$ | $\langle 5,c,17H\rangle$ |
| | ●●●\|● | $n-1$ | $15H$ | $16H$ | $d\,[5,1,1]$ | $\langle 5,c,16H\rangle$ |
| | ●\|●●● | $n-1$ | $16H$ | $18H$ | $d\,[5,1,1]$ | |
| | \|●●●\| | $n-2$ | $12H$ | $12H$ | $d\,[5,2,1]$ | $\langle 5,2c,12H\rangle$ |
| 6 | ●●●\|●● | $n-1$ | $18H$ | $20H$ | $d\,[6,1,1]$ | $\langle 6,c,21H\rangle$ |
| | ●●\|●●● | $n-1$ | $20H$ | $24H$ | $d\,[6,1,1]$ | |
| | ●\|●●●\| | $n-2$ | $16H$ | $18H$ | $d\,[6,2,1]$ | $\langle 6,2c,16H\rangle$ |
| | \|●●●\|● | $n-2$ | $15H$ | $16H$ | $d\,[6,2,1]$ | |
| 7 | ●●●\|●●● | $n-1$ | $24H$ | $24H$ | $d\,[7,1,2]$ | $\langle 7,c,24H\rangle$ |
| | ●●\|●●●\| | $n-2$ | $20H$ | $24H$ | $d\,[7,2,1]$ | $\langle 7,2c,21H\rangle$ |
| | ●\|●●●\|● | $n-2$ | $19H$ | $22H$ | $d\,[7,2,1]$ | |
| | \|●●●\|●● | $n-2$ | $18H$ | $20H$ | $d\,[7,2,1]$ | |

**Fig. 3.** Distances up to level 7.

by ●●, ●●, ●| and |●. Figure 3 shows the distances, bounds, and multiplicities for level 2 to 7.

Observe that $d\,[2,0,0]$ and $d\,[6,1,1]$ are in a sense "equivalent", since they are used for cases that differ only in one complete block of three $z$–pseudoatoms and one $c$–pseudoatom. Hence, we could have written $d\,[6,1,1] = d\,[2,0,0] + \langle 4,c,h\rangle$ instead. Moreover, $d\,[6,2,1] = d\,[2,1,0] + \langle 4,c,h\rangle$ and $d\,[7,2,1] = d\,[3,1,0] + \langle 4,c,h\rangle$. Similarly, distances of level greater than 7 can be decomposed into a distance of low level (4 to 7) and an appropriate number of blocks of three $z$–pseudoatoms and one $c$–pseudoatom. We set $\beta := \langle 4,c,h\rangle$ and define in Figure 4 the distances of level 8 to $4n-5$. In the table, the number of blocks $k$ varies from 1 to $n-3$. Finally, in Figure 5 the distances that have level $4n-4$ to $4n-1$ are shown. Observe that as before they are derived from distances of level 4 to 7, for $k = n-2$. However, not all combinations are necessary for these distances.

Our distance multiset $D$ consists of all atoms $z_i$ and $c_i$, and all distances specified in Figures 3, 4 and 5, with the corresponding multiplicities. There are $4n-1$ levels, and for each level $\ell$ there are $4n-\ell$ distances in $D$. In total, this yields $\sum_{\ell=1}^{4n-1}(4n-\ell) = \binom{4n}{2}$ distances. The cardinality of $D$ is polynomially bounded in $n$, and each distance in $D$ is polynomial in $h$. Hence, multiset $D$ can be constructed in polynomial time from a given instance of 3-Partition.

In parallel to the definition of $D$, we have shown already that a solution for the 3-Partition instance yields a solution for the PD–AbsError instance. In the following, we show the opposite direction, i.e., we show that a solution for the PD–AbsError instance yields a solution for the 3-Partition instance. Let $R = \{r_1,\ldots,r_{4n}\}$ be *any* set of $4n$ points on a line that is a solution for the PD–AbsError instance, i.e., multiset $D$ is the multiset of pairwise distances

| level $\ell$ | pseudoatoms | multiplicity | distance name | distance value |
|---|---|---|---|---|
| $4k+4$ | $\bullet\bullet\mid\ldots\mid\bullet$ | $n-k-1$ | $d\,[4+4k, 1+k, 0+k]$ | $d\,[4,1,0]+k\cdot\beta$ |
| | $\bullet\mid\ldots\mid\bullet\bullet$ | $n-k-1$ | $d\,[4+4k, 1+k, 0+k]$ | |
| | $\bullet\bullet\bullet\mid\ldots\mid$ | $n-k-1$ | $d\,[4+4k, 1+k, 1+k]$ | $d\,[4,1,1]+k\cdot\beta$ |
| | $\mid\ldots\mid\bullet\bullet\bullet$ | $n-k-1$ | $d\,[4+4k, 1+k, 1+k]$ | |
| $5+4k$ | $\bullet\bullet\mid\ldots\mid\bullet\bullet$ | $n-k-1$ | $d\,[5+4k, 1+k, 0+k]$ | $d\,[5,1,0]+k\cdot\beta$ |
| | $\bullet\bullet\bullet\mid\ldots\mid\bullet$ | $n-k-1$ | $d\,[5+4k, 1+k, 1+k]$ | $d\,[5,1,1]+k\cdot\beta$ |
| | $\bullet\mid\ldots\mid\bullet\bullet\bullet$ | $n-k-1$ | $d\,[5+4k, 1+k, 1+k]$ | |
| | $\mid\ldots\mid\bullet\bullet\bullet\mid$ | $n-k-2$ | $d\,[5+4k, 2+k, 1+k]$ | $d\,[5,2,1]+k\cdot\beta$ |
| $6+4k$ | $\bullet\bullet\bullet\mid\ldots\mid\bullet\bullet$ | $n-k-1$ | $d\,[6+4k, 1+k, 1+k]$ | $d\,[6,1,1]+k\cdot\beta$ |
| | $\bullet\bullet\mid\ldots\mid\bullet\bullet\bullet$ | $n-k-1$ | $d\,[6+4k, 1+k, 1+k]$ | |
| | $\bullet\mid\ldots\mid\bullet\bullet\bullet\mid$ | $n-k-2$ | $d\,[6+4k, 2+k, 1+k]$ | $d\,[6,2,1]+k\cdot\beta$ |
| | $\mid\ldots\mid\bullet\bullet\bullet\mid\bullet$ | $n-k-2$ | $d\,[6+4k, 2+k, 1+k]$ | |
| $7+4k$ | $\bullet\bullet\bullet\mid\ldots\mid\bullet\bullet\bullet$ | $n-k-1$ | $d\,[7+4k, 1+k, 2+k]$ | $d\,[7,1,2]+k\cdot\beta$ |
| | $\bullet\bullet\mid\ldots\mid\bullet\bullet\bullet\mid$ | $n-k-2$ | $d\,[7+4k, 2+k, 1+k]$ | $d\,[7,2,1]+k\cdot\beta$ |
| | $\bullet\mid\ldots\mid\bullet\bullet\bullet\mid\bullet$ | $n-k-2$ | $d\,[7+4k, 2+k, 1+k]$ | |
| | $\mid\ldots\mid\bullet\bullet\bullet\mid\bullet\bullet$ | $n-k-2$ | $d\,[7+4k, 2+k, 1+k]$ | |

**Fig. 4.** Distances with level 8 to $4n-5$ (with $\beta = \langle 4, c, h\rangle$). Value $k$ varies between 1 and $n-3$.

of $R$, up to additive error $\delta$ for each distance. We assume w.l.o.g. that the points are ordered from left to right, i.e., $r_1 < r_2 < \ldots < r_{4n}$. We will show that $R$ is basically identical to $P$, the point set that we constructed above.

Obviously, additive error $\delta$ can affect only the last digit of each distance, since base $Z$ is sufficiently large. Thus, exactly those distances with value 1 in the first digit are atoms, since all other distances have value greater than 1 in the first digit, and since there must be exactly $4n-1$ atoms. This implies immediately that the first digit of each distance denotes the level of the distance in any solution.

We now show that error $+\delta$ has to be applied to each single atom to make it fit to the distances between adjacent points in $R$. To see this, first observe that the atoms sum up to $\sum_{i=1}^{3n} z_i + \sum_{i=1}^{n-1} c_i = \langle 4n-1, (n-1)c, nh\rangle - (4n-1)\delta$. On the other hand, $d\,[4n-1, n-1, n] = \langle 4n-1, (n-1)c, nh\rangle + \delta$ is the largest distance in $D$. Each atom is the distance between two adjacent points in $R$, up to additive error $\delta$, while $d\,[4n-1, n-1, n]$ is the distance between the first and the last point in $R$, again up to additive error $\delta$. Hence, the atoms must sum up to the length of the largest distance. This is only possible if we apply error $+\delta$ to each atom, yielding sum $\langle 4n-1, (n-1)c, nh\rangle$, and if we apply error $-\delta$ to the largest distance, yielding $\langle 4n-1, (n-1)c, nh\rangle$ as well. Knowing this, we can again define *pseudoatoms* $\hat{z}_i = z_i + \delta$ and $\hat{c}_i = c_i + \delta$, which represent exactly the distances of adjacent points in $R$ (without error). Observe that if we represented the distances between adjacent points in $R$ in our number representation, then pseudoatom $\hat{z}_i$ would have exactly value $q_i$ in its last digit, for all $1 \le i \le 3n$.

We now show that the ordering of the pseudoatoms arising from $R$ is such that there are $n$ blocks of three pseudoatoms $\hat{z}_i$, and each two blocks are separated by one pseudoatom $\hat{c}_i$. Between any two adjacent $c$–pseudoatoms there must be exactly three $z$–pseudoatoms: Since there are no distances of level 4 with value $2c$ in the second digit, no combination $||$ or $|\bullet|$ or $|\bullet\bullet|$ is possible, and there are at least three $z$–pseudoatoms in between two $c$–pseudoatoms; moreover, since there

| level $\ell$ | lower bound | upper bound | distance name | distance value |
|---|---|---|---|---|
| $4n-4$ | $(n-2)h+11H$ | $(n-2)h+16H$ | $d\,[4n-4, n-1, n-2]$ | $d\,[4,1,0]+(n-2)\cdot\beta$ |
| | $(n-2)h+10H$ | $(n-2)h+14H$ | $d\,[4n-4, n-1, n-2]$ | |
| | $(n-1)h$ | $(n-1)h$ | $d\,[4n-4, n-1, n-1]$ | $d\,[4,1,1]+(n-2)\cdot\beta$ |
| | $(n-1)h$ | $(n-1)h$ | $d\,[4n-4, n-1, n-1]$ | |
| $4n-3$ | $(n-1)h+3H$ | $(n-1)h+4H$ | $d\,[4n-3, n-1, n-1]$ | $d\,[5,1,1]+(n-2)\cdot\beta$ |
| | $(n-1)h+4H$ | $(n-1)h+6H$ | $d\,[4n-3, n-1, n-1]$ | |
| | $(n-2)h+14H$ | $(n-2)h+20H$ | $d\,[4n-3, n-1, n-2]$ | $d\,[5,1,0]+(n-2)\cdot\beta$ |
| $4n-2$ | $(n-1)h+6H$ | $(n-1)h+8H$ | $d\,[4n-2, n-1, n-1]$ | $d\,[6,1,1]+(n-2)\cdot\beta$ |
| | $(n-1)h+8H$ | $(n-1)h+12H$ | $d\,[4n-2, n-1, n-1]$ | |
| $4n-1$ | $nh$ | $nh$ | $d\,[4n-1, n-1, n]$ | $\langle 4n-1, (n-1)c, nh\rangle+\delta$ |

**Fig. 5.** Distances with level $4n-4$ to $4n-1$. Each case occurs once.

are $n-2$ distances of level 5 with value $2c$ in the second digit, there must be at least $n-1$ $c$–pseudoatoms such that there are always at most 3 $z$–pseudoatoms in between. Hence, the points in $R$ are such that blocks of three $z$–pseudoatoms alternate with one $c$–pseudoatom, starting and ending with a block of three $z$–pseudoatoms.

Finally, we show that the third digits of each three adjacent $z$–pseudoatoms sum up to $h$: Consider those distances of level 3 that have a zero in the second digit. There are $n$ such distances, and their third digits sum up to $nh+n\delta$. Each of these distances must span exactly one of the $n$ blocks of three $z$–pseudoatoms. The total sum of the last digit of all $z$–pseudoatoms is exactly $\sum_{i=1}^{3n} q_i = nh$. Since the distances of level 3 that span these blocks do not overlap, they have to sum up to the same total. Hence, the error for each such distance of level 3 must be $-\delta$. This implies that each three $q_i$'s that correspond to one block sum up to exactly $h$ (since we have applied error $+\delta$ to each atom to define the $z$–pseudoatoms). Thus, these triples yield a solution for the 3-Partition instance. □

## 3 Strong NP-completeness of PD–RelError

In this section, we show that PD–RelError is strongly NP-complete by using a reduction from 3-Partition similar to the one used to prove strong NP-completeness of PD–AbsError (see Theorem 4).

**Theorem 5.** PD–RelError *is strongly* NP*-complete, even if the error is a constant.*

*Proof (sketch).* The problem is in NP analogously to the proof of Theorem 4. The proof of NP-hardness is also along the lines of the proof of Theorem 4. In fact, the proof has a similar structure overall, but the details are quite different. Given an instance of 3-Partition, we define a multiset $E$ of distances which are expressed as numbers with a base $Z$, with $Z = 10hnc$ and $c = n^2h^2$.

We replace the definition of the atoms as follows: $z_i = \langle 1, 0, q_i\rangle \cdot \frac{1}{1+\varepsilon}$, for $1 \le i \le 3n$, and $c_i = \langle 1, c, 0\rangle \cdot \frac{1}{1+\varepsilon}$, for $1 \le i \le n-1$. All $z_i$'s and $c_i$'s are part of the distance set $E$. Note that for a fixed level $\ell$, the corresponding distances

$d[\ell,\cdot,\cdot]$ from the proof of Theorem 4 are defined for at most two consecutive values of the second digit, say $j$ and $j+1$. Here, we define distances $e[\ell,j]$ and $e[\ell,j+1]$ for all levels $2 \le \ell \le 4n-1$ and corresponding $j$ or $j+1$, respectively, as follows: $e[\ell,j] = \langle \ell, j, B_u(\ell,j)\rangle \cdot \frac{1}{1+\varepsilon}$, and $e[\ell,j+1] = \langle \ell, j+1, B_l(\ell,j+1)\rangle \cdot \frac{1}{1-\varepsilon}$, using values $B_u()$ and $B_l()$ as specified below.

The first digit $\ell$ still indicates the level of the distance (i.e., how many atoms it will span in a solution) and the second digit $j$ or $j+1$ indicates the number of $c$-atoms it will span. Value $B_u(\ell,j)$ is the maximum upper bound from the corresponding column in Figure 3, Figure 4, or Figure 5, taken over all distances $d[\ell,j,\cdot]$ (for Figure 4, these bounds result from Figure 3 by adding appropriate multiples of $h$); similarly, value $B_l(\ell,j+1)$ is the minimum lower bound from the corresponding column in the figures, taken over all distances $d[\ell,j+1,\cdot]$. The multiplicity of distance $e[\ell,j]$ is the sum of the multiplicities for all distance values $d[\ell,j,\cdot]$ taken from the same figures, likewise for distance $e[\ell,j+1]$. Thus, for example $e[5,1] = \langle 5,1,20H\rangle \cdot \frac{1}{1+\varepsilon}$ with multiplicity $3(n-1)$, while $e[6,2] = \langle 6,2,15H\rangle \cdot \frac{1}{1-\varepsilon}$ with multiplicity $2(n-2)$.

For $d[\cdot]$-distances with levels divisible by four (i.e., distances $d[4\ell',j,\cdot]$ with integer $\ell' < n$), we only have one possible value $j$ for the second digit. Thus, we define the corresponding $e[\cdot]$-distances by $e[4\ell',j] = \langle 4\ell', j, B_u(4\ell',j)\rangle \cdot \frac{1}{1-\varepsilon}$. Finally, we define two special distances: $e[3,0] = \langle 3,0,h\rangle \cdot \frac{1}{1+\varepsilon}$, with multiplicity $n$, and $e[4n-1,n-1] = \langle 4n-1,(n-1)c,nh\rangle \cdot \frac{1}{1-\varepsilon}$ with multiplicity 1.

All the distances, including the atoms, are put into distance multiset $E$. We set error $\varepsilon = \frac{1}{100}$. This completes our description of how to construct a PD–RelError instance from a given 3-Partition instance. The proof that a solution for the 3-Partition instance yields a solution for the PD–RelError instance, and vice versa, as well as the strategy to transform these distances into integer distances, can be found in the full version of this paper. □

## 4 Conclusion

We have shown that Partial Digest is NP-complete if all measurements are prone to the same additive or multiplicative error. This answers the question whether Partial Digest on real-life data can be solved in polynomial time. However, it also gives rise to new questions: While we have shown NP-hardness for even constant relative error, our proof for absolute error uses error $\frac{h}{4}$, which is not constant. Is Partial Digest still NP-complete if we restrict the additive error to some (small) constant? What if we allow only one-sided errors (i.e., if the lengths of the distances are always underestimated)? Moreover, the main open problem is still the computational complexity of Partial Digest itself.

*Acknowledgments* We would like to thank Claudio Gutiérrez, Fabian Hennecke, Roland Ulber, Birgitta Weber, and Peter Widmayer for helpful discussions, and Riko Jacob, who suggested the graphical presentation used in Section 2.

## References


1. L. Allison and C. N. Yee. Restriction site mapping is in separation theory. *Computer Applications in the Biosciences*, 4(1):97–101, 1988.
2. J. Błażewicz, P. Formanowicz, M. Kasprzak, M. Jaroszewski, and W. T. Markiewicz. Construction of DNA restriction maps based on a simplified experiment. *Bioinformatics*, 17(5):398–404, 2001.
3. M. Cieliebak, S. Eidenbenz, and P. Penna. Noisy data make the partial digest problem NP-hard. In *Proc. of the* $3^{rd}$ *Workshop on Algorithms in Bioinformatics (WABI 2003)*, pages 111–123, 2003.
4. T Dakić. *On the Turnpike Problem*. PhD thesis, Simon Fraser University, 2000.
5. T. I. Dix and D. H. Kieronska. Errors between sites in restriction site mapping. *Computer Applications in the Biosciences*, 4(1):117–123, 1988.
6. J. Fütterer. Personal communication, 2002. ETH Zurich, Institute of Plant Sciences.
7. M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of* NP*-Completeness*. Freeman, 1979.
8. J. Inglehart and P. C. Nelson. On the limitations of automated restriction mapping. *Computer Applications in the Biosciences*, 10(3):249–261, 1994.
9. P. Lemke, S. S. Skiena, and W. Smith. Reconstructing sets from interpoint distances. Technical Report TR2002–37, DIMACS, 2002.
10. P. Lemke and M. Werman. On the complexity of inverting the autocorrelation function of a finite integer sequence, and the problem of locating n points on a line, given the $\binom{n}{2}$ unlabelled distances between them. Preprint 453, Institute for Mathematics and its Application IMA, 1988.
11. G. Pandurangan and H. Ramesh. The restriction mapping problem revisited. *Journal of Computer and System Sciences*, 65(3):526–544, 2002. Special issue on Computational Biology.
12. P. A. Pevzner. *Computational Molecular Biology: An Algorithmic Approach*. MIT Press, 2000.
13. J. Rosenblatt and P. Seymour. The structure of homometric sets. *SIAM Journal of Algorithms and Discrete Mathematics*, 3(3):343–350, 1982.
14. D. B. Searls. Formal grammars for intermolecular structure. In *Proc. of the* $1^{st}$ *International Symposium on Intelligence in Neural and Biological Systems (INBS'95)*, pages 30–37, 1995.
15. J. Setubal and J. Meidanis. *Introduction to Computational Molecular Biology*. PWS Boston, 1997.
16. S. S. Skiena, W. Smith, and P. Lemke. Reconstructing sets from interpoint distances. In *Proc. of the* $6^{th}$ *ACM Symposium on Computational Geometry (SoCG 1990)*, pages 332–339, 1990.
17. S. S. Skiena and G. Sundaram. A partial digest approach to restriction site mapping. *Bulletin of Mathematical Biology*, 56:275–294, 1994.
18. P. Tuffery, P. Dessen, C. Mugnier, and S. Hazout. Restriction map construction using a 'complete sentence compatibility' algorithm. *Computer Applications in the Biosciences*, 4(1):103–110, 1988.
19. M. S. Waterman. *Introduction to Computational Biology*. Chapman & Hall, 1995.
20. Z. Zhang. An exponential example for a partial digest mapping algorithm. *Journal of Computational Biology*, 1(3):235–239, 1994.