Available online at www.sciencedirect.com

SCIENCE DIRECT®

Theoretical Computer Science 349 (2005) 361 – 381

Theoretical Computer Science

www.elsevier.com/locate/tcs

Partial Digest is hard to solve for erroneous input data

Mark Cieliebak[a], Stephan Eidenbenz[b,*], Paolo Penna[c]

[a]*Institute of Theoretical Computer Science, ETH Zurich, 8092 Zurich, Switzerland*
[b]*Discrete Simulation Sciences (CCS-5), Los Alamos National Laboratory*[1]*, USA*
[c]*Dipartimento di Informatica ed Applicazioni "Renato M. Capocelli" Università di Salerno, Italy*

Received 20 January 2004; received in revised form 27 July 2005; accepted 29 August 2005

Communicated by M. Crochemore

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 Partial Digest variants that model three different error types that can occur in the data: additional distances, missing distances, and erroneous fragment lengths. We show that these variations are NP-hard, hard to approximate, and strongly NP-hard, respectively.

Keywords: Partial Digest; Turnpike problem; Point reconstruction; Physical DNA mapping

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 20 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[2] 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 1.1 (*Partial Digest*). Given an integer m and a multiset D of $k = \binom{m}{2}$ positive integers, is there a set $P = \{p_1, \dots, p_m\}$ of m points on a line such that $\Delta(P) := \{|p_i - p_j| \mid 1 \leqslant i < j \leqslant m\} = D$?

[*] Corresponding author. Tel.: +1 505 667 3742; fax: +1 505 665 7464.
E-mail addresses: cieliebak@inf.ethz.ch (M. Cieliebak), eidenben@lanl.gov (S. Eidenbenz), penna@dia.unisa.it (P. Penna).

[1] Los Alamos National Laboratory Publication No. LA-UR-04:0361.

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

0304-3975/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2005.08.030

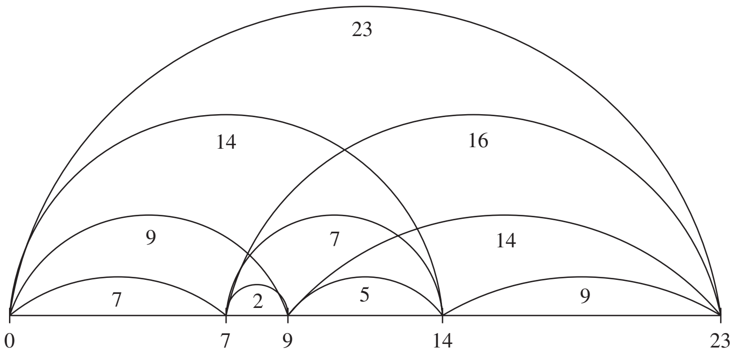

Fig. 1. Example for the Partial Digest problem.

For example, for the distance multiset $D = \{2, 5, 7, 7, 9, 9, 14, 14, 16, 23\}$, the point set $P = \{0, 7, 9, 14, 23\}$ is a feasible solution, which is shown in Fig. 1 (there exist more solutions).

1.1. Previous work

The exact computational complexity of Partial Digest 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 [29]). The problem can be solved in pseudo-polynomial time [20,26], and there exists a backtracking algorithm (for exact or erroneous data) that has expected running time polynomial in the number of distances [29,30], but exponential worst case running time [35]. The Partial Digest problem can be formalized by cut grammars, which have one additional symbol δ, the *cut*, that is neither a non-terminal nor a terminal symbol [27], and the problem is closely related to the theory of homometric sets[3] [29]. 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 [29]. However, for the original Partial Digest problem, neither a polynomial-time algorithm nor a proof of NP-hardness is known [5,10,23–25,28].

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

The Partial Digest problem occurs as well in the realm of de novo peptide sequencing using tandem mass spectrometry: given a probe with many copies of a single protein, we first use an enzyme like trypsin to cleave the proteins. This leaves us with a set of protein fragments, called peptides. We separate these peptides by their mass, and break up each single peptide into even smaller fragments using collision induced dissociation (CID). For each peptide, the mass/charge ratios of the resulting fragments are measured using mass spectrometry, yielding the *tandem mass spectrum* of the peptide. In the dissociation step, each single peptide can break up between any two amino acids in the peptide. If each single peptide breaks up exactly once, then only fragments occur that are prefixes and suffixes of the peptide sequence (e.g., peptide `AEKGCWTR` breaks up into two fragments `A` and `EKGCWRT`, or into fragments `AE` and `KGCWTR`, and so on). In this case, there exist efficient algorithms to determine the amino acid sequence of the peptide (*de novo sequencing*) [6,23]. However, in real-life experiments a single peptide does not only break up once, but it can break

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

up several times, yielding internal fragments as well [3,4,18]. In the example, peptide `AEKGCWTR` might break up into three fragments `AEK, GC` and `WTR`. In the extreme, we can obtain not only prefixes and suffixes, but fragments for *all* possible substrings of the peptide sequence. In this case, the problem to find the appropriate sequence of amino acid fragment masses is the Partial Digest problem with the additional restriction that the solution $P = \{p_1, \ldots, p_m\}$ must have consecutive points such that $|p_i - p_{i+1}|$ equals to some amino acid mass.

In the remainder of this paper, we will focus on data from partial digestion experiments; however, our results apply analogously to MS/MS spectra, too.

Due to its importance in molecular biological, many experimental variations of Partial Digest have been studied: probed partial digestion, where probes (markers) are hybridized to partially digested DNA [1,22]; simplified partial digestion, where clones are cleaved either in one or in all restriction sites [5]; labeled partial digestion, where both ends of the DNA molecule are labeled before digestion [23]; double digestion, where two different enzymes are used in complete digestion experiments to obtain three sets of distances [15]; and multiple complete digestion, where many different enzymes are used [12] (which is a generalization of double digestion). For an introduction to the Partial Digest problem, see for instance the survey by Lemke et al. [19], and the books by Pevzner [24] or by Setubal and Meidanis [28].

1.3. Erroneous data

In reality, the partial digest experiment cannot be conducted under ideal conditions, and thus errors occur in the data. In fact, there are four types of errors that occur in partial digest experiments [11,13,17,30,34]:

Additional fragments. An enzyme may erroneously cut in some cases at a site that is similar, but not exactly equivalent to a restriction site; thus, some distances will be added to the data even though they do not belong there. Furthermore, fragments can be added through contamination with biological material, such as DNA from unrelated sources.

Missing fragments. If the enzyme fails to cut at a restriction site where it would be supposed to cut (*partial cleavage error*), this leads to missing fragments. Furthermore, fragments are not detected by gel electrophoresis if their amount is insufficient to be detected by common staining techniques. Finally, small fragments may remain undetected at all since they run off at the end of the gel.

Fragment length. Using gel electrophoresis, it is almost impossible to determine the exact length of a fragment. Typical error ranges are between 2% and 7% of the fragment length.

Multiplicity detection. Determining the proper multiplicity of a distance from the intensity of its spot in the gel is almost impossible in practice.

In this work, we will focus on the first three types of errors, and prove hardness results for each of these variations. Intuitively, the problem of modeling real-life instances—in which *all* error types can occur—is even harder than having only one error type.

1.4. Missing distances

The Min Partial Digest Superset problem models the situation of omissions, where we are given data for which we know that some distances are missing, and we search for a set of points on a line such that the number of missing distances is minimum. This problem is formally defined as follows (recall that $\Delta(P)$ denotes the multiset of all distances between any two points in P).

Definition 1.2 (*Min Partial Digest Superset*). Given a multiset D of k positive integers, find the minimum number m such that there is a set P of m points on a line with $D \subseteq \Delta(P)$.

For example, if $D = \{2, 5, 7, 7, 9, 14, 23\}$, then the point set $P = \{0, 7, 9, 14, 23\}$ (as shown in Fig. 1 on page 362) would be a minimum solution for the Min Partial Digest Superset instance D. On the other hand, if $D' = \{2, 7, 9, 9, 16\}$, then the points in P would still cover all distances from D', but there exist solutions with fewer points that cover D', e.g. point set $P' = \{0, 2, 9, 18\}$ (yielding distance multiset $\{2, 7, 9, 9, 16, 18\}$).

We show in Section 2 that computing an optimal solution for the Min Partial Digest Superset problem is NP-hard, by giving a reduction from Equal Sum Subsets. Our result provides a partial answer to the open problem 12.116 in the book by Pevzner [24], which asks for an algorithm to reconstruct a set of points, given a subset of their pairwise distances.

We can even strengthen our hardness result by considering the problem t-Partial Digest Superset, where we restrict the cardinality of a solution to at most t, for some parameter t that is specified as a fixed function in $|D|$, the cardinality of the input distance multiset:

Definition 1.3 (*t-Partial Digest Superset*). Given a multiset D of positive integers, is there a set P of $m \leqslant t$ integers such that $D \subseteq \Delta(P)$?

We show that the t-Partial Digest Superset problem is NP-hard for *any* parameter $t = f(|D|) := |D|^{(1/2)+\varepsilon}$, for any $0 < \varepsilon < \frac{1}{2}$. This result is tight in a sense, since any solution (even for the original Partial Digest) must have at least cardinality $\Omega(|D|^{1/2})$.

1.5. Additional distances

In Section 3, we study the Max Partial Digest Subset problem, which models the situation of additions: we are given data in which some wrong distances were added, and we search for a set of points on a line such that they cover a maximum number of the given distances. A formal definition is as follows.

Definition 1.4 (*Max Partial Digest Subset*). Given a multiset D of k positive integers, find the maximum number m such that there is a set P of m points on a line with $\Delta(P) \subseteq D$.

We show that there is no polynomial-time algorithm for this problem that guarantees an approximation ratio of $|D|^{(1/2)-\varepsilon}$ for any $\varepsilon > 0$, unless $\mathsf{NP} = \mathsf{ZPP}$.[4] To establish this result, we give a gap-preserving reduction from Max Clique. We also point to a trivial approximation algorithm for Max Partial Digest Subset that achieves a matching asymptotic approximation ratio. Thus, our inapproximability result is tight up to low-order terms.

Our two optimization variations of the Partial Digest problem allow the multiset of pairwise distances in a solution to be either a superset (i.e., to cover all given distances in D plus additional ones) or a subset (i.e., to contain only some of the distances in D) of the input set D. If a polynomial-time algorithm existed for either Min Partial Digest Superset or Max Partial Digest Subset, we could use this algorithm to solve the original Partial Digest problem as well: any YES instance of Partial Digest is an instance of both optimization problems whose optimum is $\frac{1}{2} + \sqrt{\frac{1}{4} + 2k}$; any NO instance of Partial Digest is an instance of Max Partial Digest Subset (resp., Min Partial Digest Superset) whose optimum is at most $\frac{1}{2} + \sqrt{\frac{1}{4} + 2k} - 1$ (at least $\frac{1}{2} + \sqrt{\frac{1}{4} + 2k} + 1$, respectively).

1.6. Fragment length measurement errors

As a third type of error that can occur in real-life data, we study Partial Digest with inaccurate distance lengths. In Sections 4 and 5, we will show that measurement errors make the Partial Digest problem strongly NP-complete, for both additive or multiplicative errors.

Algorithms for Partial Digest with inaccurate data have been studied intensively in the literature [11,17,30], and different error models have been designed, e.g. for measurement errors that are logarithmic in the size of the fragment length [31,32] or for intervals of absolute errors [2,30].

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* [19,29]. 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, and moreover, we cannot expect individual error bounds. Therefore, we study the computational complexity of the variation of Partial Digest where *all* measurements are prone to *the same additive non-zero* error.

[4] A problem Π is in class ZPP if there is a probabilistic algorithm for Π with polynomial running time which never outputs a wrong result, and which fails with probability less than $\frac{1}{2}$.

We say that value v matches a distance d up to (additive) error ε if $|v - d| \leqslant \varepsilon$; moreover, a multiset D is a distance multiset for point set P up to error ε, if each distance between any two points in P can be matched with a distance in D up to error ε, and this matching is bijective. The Partial Digest with Additive Errors problem is defined as follows.

Definition 1.5 (*Partial Digest with Additive Errors*). Given an integer m, a multiset D of $k = \binom{m}{2}$ positive integers, and an error bound $\varepsilon > 0$, is there a set P of m points on a line such that D is the distance multiset for P up to error ε?

In Section 4, we prove that Partial Digest with Additive Errors 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* $\mathrm{e} > 0$ if $d(1 - \mathrm{e}) \leqslant x \leqslant d(1 + \mathrm{e})$. This problem variant models the situation where measurement errors can be bounded by a percentage of the fragment lengths. Observe that this definition is not symmetric, i.e., if d matches x up to error e, 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 e 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 e. The Partial Digest with Relative Error problem is defined as follows.

Definition 1.6 (*Partial Digest with Relative Error*). Given an integer m, a multiset D of $k = \binom{m}{2}$ positive integers, and a rational error $\mathrm{e} > 0$, is there a set P of m points on a line such that D is the distance multiset of P up to multiplicative error e?

We show in Section 5 that Partial Digest with Relative Error is strongly NP-complete, even for constant error, by using a similar reduction as for Partial Digest with Additive Errors.

1.7. Notation

We introduce a vector representation for large numbers that will allow us to add up the numbers digit by digit, like polyadic numbers. The numbers are expressed in the number system of some base Z. We denote by $\langle a_1, \ldots, a_n \rangle$ the number $\sum_{1 \leqslant i \leqslant n} a_i Z^{n-i}$; we say that a_i is the ith 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 $Z = 27$ and numbers $\alpha = \langle 3, 5, 1 \rangle$, $\beta = \langle 2, 1, 0 \rangle$, then $\alpha + \beta = \langle 5, 6, 1 \rangle$ and $3 \cdot \alpha = \langle 9, 15, 3 \rangle$. We define the concatenation of two numbers by $\langle a_1, \ldots, a_n \rangle \circ \langle b_1, \ldots, b_m \rangle := \langle a_1, \ldots, a_n, b_1, \ldots, b_m \rangle$, i.e., $\alpha \circ \beta = \alpha Z^m + \beta$, where m is the number of digits in β. Let $\Delta_n(i) := \langle 0, \ldots, 0, 1, 0, \ldots, 0 \rangle$ be the number that has n digits, all 0's except for the ith position, where the digit is 1. Moreover, $\mathbb{1}_n := \langle 1, \ldots, 1 \rangle$ has n digits, all 1's, and $\mathbb{0}_n := \langle 0, \ldots, 0 \rangle$ has n zeros. Notice that $\mathbb{1}_n = Z^n - 1$.

2. NP-hardness of Min Partial Digest Superset

In this section, we study the Min Partial Digest Superset problem and show that this problem is NP-hard by giving a reduction from Equal Sum Subsets.

First, observe that the minimum cardinality of a point set that covers all distances in a given multiset D cannot be too large. To see this, let $D = \{d_1, \ldots, d_k\}$ be a distance multiset. If m is the minimum number such that a set P of cardinality m with $D \subseteq \Delta(P)$ exists, then $m \leqslant k + 1$: We set $p_0 = 0$, $p_i = p_{i-1} + d_i$ for $1 \leqslant i \leqslant k$, and $P_{\text{triv}} = \{p_0, \ldots, p_k\}$, i.e., we simply put all distances from D in a chain "one after the other" (see Fig. 2). In P_{triv}, each distance d_i induces a new point, and we use one additional starting point 0. Obviously, set P_{triv} covers D and has cardinality $k + 1$.

Observe that Partial Digest can be easily reduced to Min Partial Digest Superset: given an instance D of Partial Digest of cardinality $|D| = k$, there is a solution for D if and only if the minimal solution for the Min Partial Digest Superset instance D has size $m = \frac{1}{2} + \sqrt{\frac{1}{4} + 2k}$ $\left(\text{in this case, } k = \binom{m}{2}\right)$.

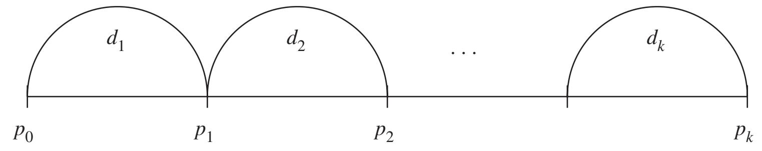

Fig. 2. Trivial solution for a distance multiset D.

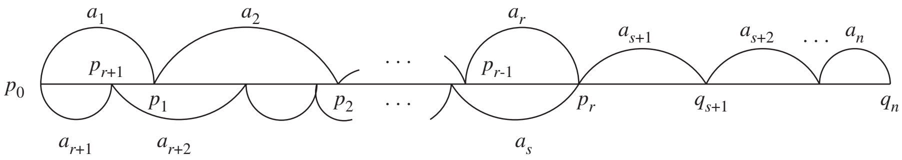

Fig. 3. Solution if there are two sets of equal sum.

We now show that Min Partial Digest Superset is NP-hard by giving a reduction from Equal Sum Subsets, which is the NP-complete problem [33] that is defined as follows: given a set A of n positive integers, are there two disjoint non-empty subsets $X, Y \subseteq A$ such that $\text{sum}(X) = \text{sum}(Y)$?

Theorem 2.1. Min Partial Digest Superset *is* NP-*hard.*

Proof. We reduce Equal Sum Subsets to Min Partial Digest Superset. Given an instance $A = \{a_1, \ldots, a_n\}$ of Equal Sum Subsets, we set $D = A$ (and $k = n$), and prove in the following that there is a solution for the Equal Sum Subsets instance A if and only if a minimal solution for the Min Partial Digest Superset instance D has at most n points.

Let X and Y be a solution for the Equal Sum Subsets instance. Assume w.l.o.g. that $X = \{a_1, \ldots, a_r\}$ and $Y = \{a_{r+1}, \ldots, a_s\}$, for some $1 \leqslant r < s \leqslant n$. We construct a set P that covers D and that has at most cardinality n. Similarly to the construction of P_{triv}, we line up the distances from D. In this case, *two* chains start at point 0: those distances from X and those from Y (see Fig. 3); the remaining distances from $D - (X \cup Y)$ are positioned at the end of the two chains. More precisely, we set

$$\begin{aligned}
&p_0 = 0\\
&p_i = p_{i-1} + a_i \quad \text{for } 1 \leqslant i \leqslant r\\
&p_{r+1} = a_{r+1}\\
&p_j = p_{j-1} + a_j \quad \text{for } r+2 \leqslant j \leqslant s-1\\
&q_{s+1} = p_r + a_{s+1}\\
&q_\ell = q_{\ell-1} + a_\ell \quad \text{for } s+2 \leqslant \ell \leqslant n.
\end{aligned}$$

Set $P = \{p_0, \ldots, p_{s-1}, q_{s+1}, \ldots, q_n\}$ is the corresponding set of points. Notice that there is no point "p_s" in set P, since the two chains corresponding to X and Y share two points, namely $p_0 = 0$ and their common endpoint p_r.

Obviously, P is a set of cardinality n. Moreover, the definition of the points yields immediately that except for $i = s$ each a_i is the difference between two of the points (either $p_i - p_{i-1}$, or $q_{s+1} - p_r$, or $q_\ell - q_{\ell-1}$). To see that a_s occurs as well, first observe that $p_r = \sum_{i=1}^{r} a_i = \text{sum}(X)$ and that $p_{s-1} = \sum_{j=r+1}^{s-1} a_j = \text{sum}(Y) - a_s$. Thus, $p_r - p_{s-1} = \text{sum}(X) - (\text{sum}(Y) - a_s) = a_s$, since X and Y are a solution of the Equal Sum Subsets instance and yield the same sum. Hence, P covers every distance from D.

For the opposite direction, let $P = \{p_1, \ldots, p_m\}$ be an optimal solution for the Min Partial Digest Superset instance with $m < n + 1$. Since P covers D, for each $a \in D$ there is a pair (p, q) of points $p, q \in P$ such that $a = |p - q|$. For each $a \in D$ we choose one such pair and say that it is *associated* with value a. We define a graph $G = (V, E)$ with $V = P$ and $E = \{(p, q) \mid (p, q) \text{ is associated with some } a \in D\}$, i.e., G contains only those edges corresponding to some distance in D. Thus, $|V| = m$ and $|E| = |D| = n$. Since $m < n + 1$, this graph contains at least one cycle. We show in the following that such a cycle induces a solution for the Equal Sum Subsets instance.

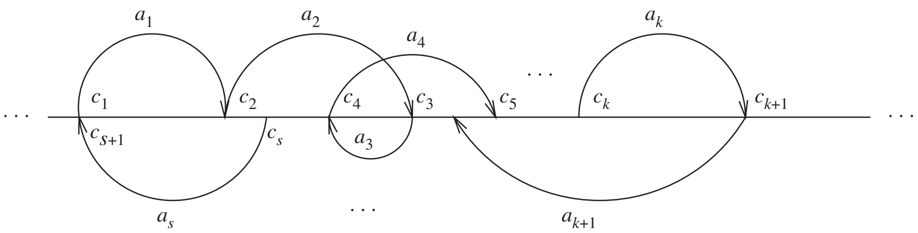

Fig. 4. A solution containing a cycle yields two subsets of equal sum: the overall length of right jumps equals the overall length of left jumps.

Let $C = c_1, \dots, c_s$ be any cycle in G (see Fig. 4). Then $|c_{i+1} - c_i| \in D$, for all $1 \leqslant i \leqslant s$ (here, we abuse notation and identify c_{s+1} with c_1). Assume w.l.o.g. that $|c_{i+1} - c_i|$ is associated with a_i, for $1 \leqslant i \leqslant s$. We define $I^+ := \{i \in \{1, \dots, s\} \mid c_{i+1} > c_i\}$, and $I^- := \{j \in \{1, \dots, s\} \mid c_{j+1} < c_j\}$, i.e., we partition the edges in the cycle into two sets, those that are oriented to the left (I^-) and those that are oriented to the right (I^+). This yields

$$\begin{aligned} 0 &= c_{s+1} - c_1 \\ &= \sum_{i=1}^{s} (c_{i+1} - c_i) \\ &= \sum_{i \in I^+} (c_{i+1} - c_i) + \sum_{j \in I^-} (c_{j+1} - c_j) \\ &= \sum_{i \in I^+} |c_{i+1} - c_i| - \sum_{j \in I^-} |c_{j+1} - c_j| \\ &= \sum_{i \in I^+} a_i - \sum_{j \in I^-} a_j. \end{aligned}$$

Sets $X := \{a_i \mid i \in I^+\}$ and $Y := \{a_j \mid j \in I^-\}$ yield equal sums, and thus a solution for the Equal Sum Subsets instance. □

In the previous theorem, we have shown NP-hardness of Min Partial Digest Superset by reduction from Equal Sum Subsets. In the proof, we distinguished whether a minimal solution uses at most n points, or $n + 1$ points (which in fact are always sufficient). We will now extend this result and allow to "decrease" the bound to some value t that is still sufficiently large. In fact, we show that the corresponding problem t-Partial Digest Superset is NP-hard for every $0 < \varepsilon < \frac{1}{2}$, if we set t to be $f(|D|) = |D|^{(1/2)+\varepsilon}$. Observe that for a distance multiset D, a minimal set of points covering D has cardinality at least $\frac{1}{2} + \sqrt{\frac{1}{4} + 2|D|} \approx |D|^{1/2}$. Moreover, the Partial Digest problem is equivalent to t-Partial Digest Superset with $t = \frac{1}{2} + \sqrt{\frac{1}{4} + 2|D|} = \mathrm{O}(|D|^{1/2})$.

Theorem 2.2. *t-Partial Digest Superset is NP-hard for any constant $0 < \varepsilon < \frac{1}{2}$ and for any $t = f(|D|) := |D|^{(1/2)+\varepsilon}$.*

Proof. We show NP-hardness by reduction from Equal Sum Subsets, analogous to the proof of Theorem 2.1. Let $\{a_1, \dots, a_n\}$ be an instance of Equal Sum Subsets. Informally speaking, we "blow up" the instance of Min Partial Digest Superset used in the proof of Theorem 2.1 (cf. Fig. 5): First, we have n distances in a set A', each corresponding to one of the a_i's. Then we add a set B of q "essential" distances (for some value q that we specify later) such that any solution for our instance must use exactly $q + 1$ points to cover the distances in B, and no two of these points can be used to cover any distances from A'. Finally, we add a suitable set C' of $\mathrm{O}(q^2)$ "inessential" distances to fill up the number of distances in our instance. Each distance in C' is the sum of some distances from B, and all the distances in C' can be covered "for free" by the points used for the distances in B (i.e., no additional points are necessary). Our instance D for t-Partial Digest Superset is the union of the distance sets A', B and C'. We will choose the size of set C' such that $t = f(|D|) = n + q$ holds. Moreover, we will show that either $n + q$ points are sufficient to cover all distances in our instance, or that we need at least $n + q + 1$ points, and that there is a solution for the Equal Sum Subsets instance if and only if $n + q$ points are sufficient.

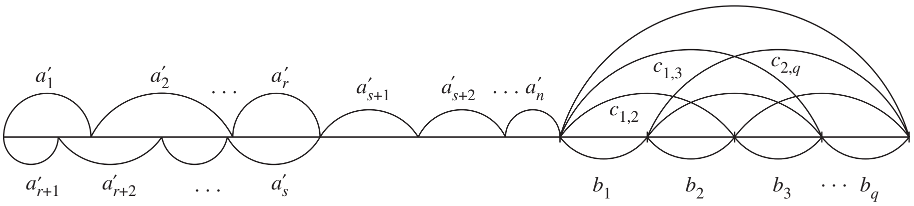

Fig. 5. Distance sets A', B and C.

We postpone the choice of q and show first how the distance sets are defined. All distances are numbers with base $Z = q^2 + \sum_{i=1}^{n} a_i$. Let $a'_i = \langle a_i \rangle \circ \mathbb{0}_q$ and $A' = \{a'_i \mid 1 \leqslant i \leqslant n\}$. For $1 \leqslant j \leqslant q$, let $b_j = \langle 0 \rangle \circ \Delta_q(j)$, and $B = \{b_j \mid 1 \leqslant j \leqslant q\}$. For $1 \leqslant u < v \leqslant q$, let $c_{u,v} = \sum_{\ell=u}^{v} b_\ell$, and $C = \{c_{u,v} \mid 1 \leqslant u < v \leqslant q\}$. Obviously, no distances from A' sum up to a distance in B or C, and vice versa.

The instance of t-Partial Digest Superset is defined by $D = A' \cup B \cup C'$, where C' is a subset of C of appropriate size. Clearly, $|D| = n + q + |C'|$. We want to choose the size of $|C'|$ such that $f(|D|) = (n + q + |C'|)^{(1/2)+\varepsilon} = n + q$ is satisfied. To this end, it suffices to take any $C' \subseteq C$ with cardinality $|C'| = (n+q)^{2/(1+2\varepsilon)} - (n+q)$. (If the latter number is not an integer, the proof can be easily adjusted by considering $|C'| = \lfloor (n+q)^{2/(1+2\varepsilon)} \rfloor - (n+q)$, and choosing q appropriately; this is possible for sufficiently large n). In order to make this possible, we need to have $|C| \geqslant (n+q)^{2/(1+2\varepsilon)} - (n+q)$. Since C contains $\binom{q}{2}$ distances, we have to choose q sufficiently large to make the inequality $\binom{q}{2} \geqslant (n+q)^{2/(1+2\varepsilon)} - (n+q)$ hold. This inequality holds if we choose $q \geqslant \max\{6^{1/\varepsilon}, n\}$, which is shown as follows.

$$\begin{array}{llll}
 & q & \geqslant & 6^{1/\varepsilon} \quad \text{(by assumption)} \\
\Rightarrow & q^{1/2} & \geqslant & 6^{1/2\varepsilon} \\
\Rightarrow & q-2 & \geqslant & 6^{1/2\varepsilon} \quad (\text{since } q-2 > q^{1/2} \text{ for } q \geqslant 6^{1/\varepsilon} > 6) \\
\Rightarrow & (q-2)^{2\varepsilon} & \geqslant & 6 \\
\Rightarrow & (q-2)^{2\varepsilon} & \geqslant & \dfrac{q}{q-2} \cdot 4 \quad \left(\text{since } \dfrac{3}{2} > \dfrac{q}{q-2} \text{ for } q > 6\right) \\
\Rightarrow & (q-2)^{2\varepsilon} & \geqslant & \dfrac{q}{q-2} \cdot 2^{(3+2\varepsilon)/2} \quad (\text{since } 4 > 2^{(3+2\varepsilon)/2} \text{ for } \varepsilon < \tfrac{1}{2})
\end{array}$$

$$\begin{array}{llll}
\Rightarrow & (q-2)^{1+2\varepsilon} & \geqslant & 2q \cdot \sqrt{2}^{1+2\varepsilon} \\
\Rightarrow & \left(\dfrac{q-2}{\sqrt{2}}\right)^{1+2\varepsilon} & \geqslant & n+q \quad (\text{since } q \geqslant n) \\
\Rightarrow & \dfrac{(q-2)^2}{2} & \geqslant & (n+q)^{2/(1+2\varepsilon)} \\
\Rightarrow & \dbinom{q}{2} & \geqslant & (n+q)^{2/(1+2\varepsilon)}
\end{array}$$

We claim that there are two subsets of A of equal sum if and only if there is a set P of at most $t = n + q$ points such that $D \subseteq \Delta(P)$. The proof of this equivalence is based on the fact that, by construction, no subset of distances from $B \cup C'$ can have the same length as a subset of distances from A'. Therefore, we need $q + 1$ points to cover all distances from $B \cup C'$. The remaining set A' behaves as in the proof of Theorem 2.1: by reusing one of the $q + 1$ points, we need at most n further points to cover A'; as in the proof of Theorem 2.1, less than n points are sufficient if and only if there exists a solution for the Equal Sum Subsets instance. □

3. Approximability of Max Partial Digest Subset

In this section, we show that Max Partial Digest Subset is as hard to approximate as Max Clique, and we give a trivial approximation algorithm that achieves a matching approximation ratio.

In the following, we construct a gap-preserving reduction from Max Clique to Max Partial Digest Subset, where Max Clique is defined as follows: given a graph $G = (V, E)$ with vertices V and edges E, find a maximum clique in G, i.e., a maximum complete subgraph of G. The problem Max Clique cannot be approximated by any polynomial-time algorithm to within factor $n^{1-\varepsilon}$ for any constant $\varepsilon > 0$, where n is the number of vertices of the input graph, unless $\mathsf{NP} = \mathsf{ZPP}$ [16,21]. Our reduction is gap-preserving, thus the inapproximability of Max Clique is transferred to Max Partial Digest Subset.

Theorem 3.1. Max Partial Digest Subset *cannot be approximated to within factor* $|D|^{(1/2)-\varepsilon}$, *for any constant* $\varepsilon > 0$, *where* $|D|$ *is the number of input distances*, *unless* $\mathsf{NP} = \mathsf{ZPP}$.

Proof. Let $G = (V, E)$ be an instance of Max Clique with vertex set $V = \{v_1, \ldots, v_n\}$ and edge set $E \subseteq V \times V$. We construct an instance D of Max Partial Digest Subset by creating a number $d_{i,j} = \mathbb{0}_i \circ \mathbb{1}_{j-i} \circ \mathbb{0}_{n-j}$ with base $Z = n^2 + 1$ for each $(v_i, v_j) \in E$, $j > i$.

Let OPT be the size of a maximum clique in G (i.e., the number of vertices in a maximum clique), let OPT' be the maximum number of points that can be placed on a line such that all pairwise distances appear in D, let $k > 0$ be an integer, and let $\varepsilon > 0$. We now prove the following two implications.

(1) If $OPT \geqslant kn^{1-\varepsilon}$, then $OPT' \geqslant kn^{1-\varepsilon}$.

(2) If $OPT < k$, then $OPT' < k$.

To see the first implication, assume we are given a clique in graph G of size $kn^{1-\varepsilon}$. We construct a solution for the corresponding Max Partial Digest Subset instance D by positioning a point at position $v'_i = \mathbb{1}_i \circ \mathbb{0}_{n-i}$ for each vertex v_i in the clique. This yields a feasible solution for D, since for $j > i$ each distance $v'_j - v'_i = \mathbb{0}_i \circ \mathbb{1}_{j-i} \circ \mathbb{0}_{n-j} = d_{i,j}$ between two points v'_j and v'_i corresponds to an edge in G and is therefore encoded as distance $d_{i,j}$ in D.

We now show the second implication by proving its converse, i.e., by showing $OPT' \geqslant k \implies OPT \geqslant k$. Suppose we are given a solution of the Max Partial Digest Subset instance consisting of k points $p_1 < \cdots < p_k$ on a line. We assume w.l.o.g. that $p_1 = \mathbb{0}_n$. Let $d_{i_{\min}, j_{\max}} = p_k - p_1$. Note that $d_{i_{\min}, j_{\max}}$, and thus $i_{\min}$ and $j_{\max}$, are uniquely defined by construction. Each of the points $p_2, \ldots, p_{k-1}$ from the solution has the following properties:

(1) It only has zeros and ones in its digits, as the distance to point p_1 would not be in D otherwise.

(2) It only has zeros in the first $i_{\min}$ digits, as the distance to point p_k would not be in D otherwise.

(3) It contains at most one continuous block of ones in its digits, as the distance to point p_1 would not be in D otherwise.

The points $p_2, \ldots, p_{k-1}$ also have the property that they are of the same form,

$$\begin{array}{llllll} \text{either} & \mathbb{0}_{i_{\min}} & \circ\ \mathbb{1}_{\ell} & \circ\ \mathbb{0}_{j_{\max}-\ell-i_{\min}} & \circ\ \mathbb{0}_{n-j_{\max}} & \\ \text{or} & \mathbb{0}_{i_{\min}} & \circ\ \mathbb{0}_{\ell} & \circ\ \mathbb{1}_{j_{\max}-\ell-i_{\min}} & \circ\ \mathbb{0}_{n-j_{\max}}, & \end{array}$$

where $0 \leqslant \ell \leqslant j_{\max} - i_{\min}$. Only one of the two forms can occur in a solution, since if both forms existed, i.e., at least one point of each form existed, then the distance between points of different form would not be in D, since at least one digit would not be 0 or 1.

We now construct a vertex set V' that will turn out to be a clique. Let $v_{i_{\min}}$ and $v_{j_{\max}}$ be in vertex set V'. In addition, for each point $p_{k'}$, $2 \leqslant k' \leqslant k-1$, we have one vertex in set V': if $p_{k'}$ is of the first form, i.e., $p_{k'} = \mathbb{0}_{i_{\min}} \circ \mathbb{1}_{\ell'} \circ \mathbb{0}_{j_{\max}-\ell'-i_{\min}} \circ \mathbb{0}_{n-j_{\max}}$ for some $\ell' \in \{0, \ldots, j_{\max} - i_{\min}\}$, then we include $v_{\ell'+i_{\min}}$. Analogously, if $p_{k'}$ is of the second form, i.e., $p_{k'} = \mathbb{0}_{i_{\min}} \circ \mathbb{0}_{\ell'} \circ \mathbb{1}_{j_{\max}-\ell'-i_{\min}} \circ \mathbb{0}_{n-j_{\max}}$ for some $\ell' \in \{0, \ldots, j_{\max} - i_{\min}\}$, then we include $v_{\ell'+i_{\min}}$.

In order to see that the vertex set V' is a clique, consider the difference $p_{k'} - p_{k''}$ of any two points with $k' > k''$, where $p_{k'}$ has led to the inclusion of vertex $v_{\ell'}$ into the set and $p_{k''}$ has led to the inclusion of vertex $v_{\ell''}$ into the clique. This difference is exactly $d_{\ell',\ell''}$, and thus the edge $(v_{\ell'}, v_{\ell''})$ is in E.

The promise problem of Max Clique, in which we are promised that the size of the maximum clique in a given graph G is either at least $kn^{1-\varepsilon}$, or less than k, and we are to decide which is true, is hard to decide [16]. The two implications above show that our reduction transforms this promise problem of Max Clique into a promise problem of Max Partial Digest Subset, in which we are promised that in an optimum solution of D either at least $kn^{1-\varepsilon}$, or less than k points can be placed on a line. This promise problem of Max Partial Digest Subset is hard to decide as well, since a polynomial-time algorithm for it could be used to decide the promise problem of Max Clique. Thus, unless $\mathsf{NP} = \mathsf{ZPP}$, Max Partial Digest Subset cannot be approximated with an approximation ratio of

$$\frac{kn^{1-\varepsilon}}{k} = n^{1-\varepsilon} \geqslant |D|^{1/2-\varepsilon},$$

where $|D|$ is the number of distances in instance D. This yields the claim. □

A trivial approximation algorithm for Max Partial Digest Subset works as follows: given an instance $D = \{d_1, \dots, d_{|D|}\}$, it simply places two points at distance d_1 from each other. This approximation algorithm achieves an approximation ratio of $\mathrm{O}(|D|^{1/2})$, since any optimal solution has at most $\frac{1}{2} + \sqrt{\frac{1}{4} + 2|D|}$ points. This matches our lower bound up to lower order terms.

4. Strong NP-completeness of Partial Digest with Additive Errors

In this section, we prove that Partial Digest with Additive Errors is strongly NP-complete by giving a reduction from 3-Partition, which is defined as follows: given $3n$ positive integers $q_1, \dots, q_{3n}$ and an integer h such that $\sum_{i=1}^{3n} q_i = nh$ and $\frac{h}{4} < q_i < \frac{h}{2}$, for $i \in \{1, \dots, 3n\}$, 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 [14]. 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 $q_1, \dots, q_{3n}$ and h of 3-Partition, we define a multiset of distances D and an error $\varepsilon = \frac{h}{4}$ that form an instance of Partial Digest with Additive Errors. 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 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 error ε to "close the gap" between $\frac{h}{3}$ and the true values of the q_i's.

Theorem 4.1. Partial Digest with Additive Errors *is strongly* NP*-complete.*

Proof. The problem Partial Digest with Additive Errors is obviously in NP. To prove strong NP-hardness, we give a reduction from 3-Partition. Given an instance of 3-Partition, i.e., integers $q_1, \dots, q_{3n}$ and integer h, we define a distance multiset D and an error ε that are an instance of Partial Digest with Additive Errors. There will be a solution for this instance if and only if there is a solution for the 3-Partition instance.

The set D will contain two types of distances: atoms and non-atoms. Atoms will be those distances that, in any solution P, will correspond to consecutive points. In particular, we will have an atom z_i for each q_i, and $n - 1$ additional atoms $c_1, \dots, c_n$. The idea is to define the set of non-atoms so that *every* solution P (if any) must arrange atoms in consecutive blocks of three z_i's separated by one c_j. These blocks will correspond to triples of sum h each. In particular, solution P will be forced to approximate each atom z_i and c_i by $\hat{z_i} := z_i + \varepsilon$ and $\hat{c_i} := c_i + \varepsilon$, respectively. The set D will be such that there exists a solution

$$(q_{i_1}, q_{i_2}, q_{i_3}), (q_{i_4}, q_{i_5}, q_{i_6}), \dots, (q_{i_{3n-2}}, q_{i_{3n-1}}, q_{i_{3n}})$$

for the 3-Partition instance if and only if there exists a set P of points whose consecutive distances are

$$\hat{z}_{i_1}\hat{z}_{i_2}\hat{z}_{i_3}\hat{c}_{j_1}\hat{z}_{i_4}\hat{z}_{i_5}\hat{z}_{i_6}\hat{c}_{j_2}\cdots\hat{c}_{j_{n-1}}\hat{z}_{i_{3n-2}}\hat{z}_{i_{3n-1}}\hat{z}_{i_{3n}}.$$

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, and we show how to construct a point set P that is a solution for the Partial Digest with Additive Errors instance, i.e., P matches D up to error ε. 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 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., $q_1 \leqslant q_2 \leqslant q_3$, $q_4 \leqslant q_5 \leqslant q_6$, 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 $0 \leqslant k \leqslant n-1$. Then $q_{3k+1} \leqslant \frac{h}{3}$, 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, $\frac{h}{3} \leqslant q_{3k+3}$. With $q_{3k+1} + q_{3k+2} = h - q_{3k+3}$, we have $q_{3k+1} + q_{3k+2} \leqslant h - \frac{h}{3} = \frac{2h}{3}$. In combination with the restriction $\frac{h}{4} < q_i < \frac{h}{2}$ (from the definition of 3-Partition), this yields the following inequalities:

$$\begin{aligned} \frac{h}{4} &< q_{3k+1} \leqslant \frac{h}{3} \\ \frac{h}{4} &< q_{3k+2} < \frac{h}{2} \\ \frac{h}{3} &\leqslant q_{3k+3} < \frac{h}{2} \\ \frac{h}{2} &< q_{3k+1} + q_{3k+2} \leqslant \frac{2h}{3} \\ \frac{2h}{3} &\leqslant q_{3k+2} + q_{3k+3} < h \\ h &= q_{3k+1} + q_{3k+2} + q_{3k+3} \end{aligned} \tag{1}$$

Equivalently, we can express these inequalities using $H := \frac{h}{12}$:

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

We will use these inequalities later to derive upper and lower bounds for the error that we need to apply to our distances in order to guarantee the existence of a solution for the Partial Digest with Additive Errors 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* ℓ if it spans $\ell - 1$ further points, and we say that distance d is an *atom* if it has level 1 (see Fig. 6).

We now define our instance of Partial Digest with Additive Errors and show at the same time how to construct a solution for this instance. Let $c = n^2 \cdot h^2$. Moreover, define error $\varepsilon := 3H$. 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).

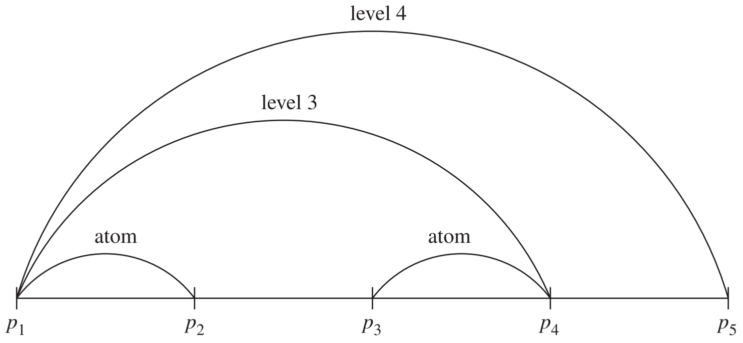

Fig. 6. Distances of different level.

First, we define $4n-1$ distances that will turn out to be atoms in our solution:

$$z_i = \langle 1, 0, q_i \rangle - \varepsilon \quad \text{for } 1 \leqslant i \leqslant 3n, \quad \text{and}$$

$$c_i = \langle 1, c, 0 \rangle - \varepsilon \quad \text{for } 1 \leqslant i \leqslant n-1.$$

Observe that the operation "$-\varepsilon$" does not affect the first digit since $\varepsilon = 3H = h/4 < q_i$ (and in fact, we could have defined z_i by $\langle 1, 0, q_i - \varepsilon \rangle$ instead) and since we choose base Z sufficiently large.

Using these distances, we can already define a "solution" P for distance multiset D (although we did not finish yet to define D; in fact, we will construct D in the following such that it matches point set P up to error ε): let $\hat{z}_i = z_i + \varepsilon$ for $1 \leqslant i \leqslant 3n$, and $\hat{c}_i = c_i + \varepsilon$ for $1 \leqslant i \leqslant n-1$. 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 Fig. 7).

We now show level by level how the distances in D are defined, and that error ε (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 error ε.

To denote the distances of higher levels we use notation $d[\ell, j, k]$, for appropriate parameters ℓ, j and k. These names already indicate the values of the three digits of a distance: distance $d[\ell, j, k]$ will have value ℓ 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 Fig. 7 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 error ε.

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 $0 \leqslant k \leqslant n-1$:

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 $\hat{c}_k$; or
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 (2) 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 ε (recall that $\varepsilon = 3H$). Observe that we could have chosen other values for the third digit of $d[2, 0, 0]$, namely any value

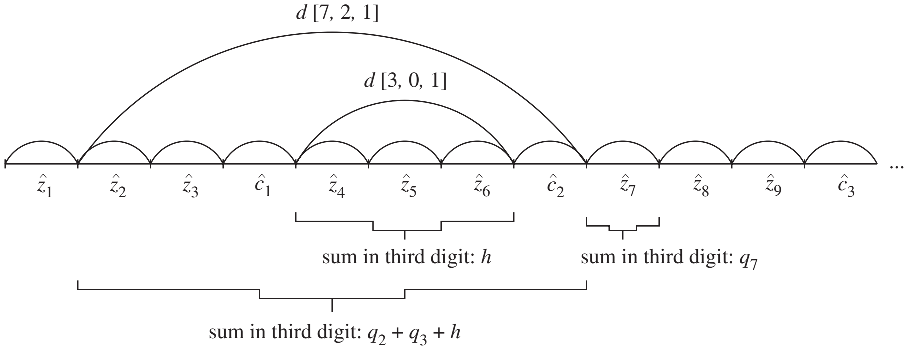

Fig. 7. Atoms and distances in multiset D.

between $5H$ and $9H$ (which still allows to match the bounds using error ε). 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 (2). 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 ε. 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 include it $2(n - 1)$ times in D, in order to cover these pairs of pseudoatoms, again up to error ε.

Before we specify the distances of higher level, we introduce a graphical representation of pseudoatoms: each z-pseudoatom is represented by a •, 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 •| • • • |, and the four cases of two adjacent pseudoatoms above can be represented by ••, ••, •| and |•.

We now define the distances of higher level. Analogously to distances of level 2, we can compute for each level the corresponding upper and lower bounds for the third digit and define appropriate distances in D. Fig. 8 shows the distances and multiplicities for level 2 to 7. This table is organized as follows. The first column specifies the level of the distance, and the second column gives the graphical representation of the combinations of pseudoatoms that can occur. The next column specifies how often each combination occurs, and the following two columns show lower and upper bounds for the third digit of the sum of the pseudoatoms. Finally, the last two columns specify the distance name that is used to cover the pseudoatoms, and the value of the distance. Distance values are only introduced once, and the lines are sorted such that those cases that use the same distance stand together.

In order to define the distances of level 8 to $4n - 2$, observe that each such distance differs from some distance of level 4, 5, 6 or 7 by a number of blocks of three z-pseudoatoms and one c-pseudoatom: e.g. distance $d[8, 2, 1]$, represented by • • | • • • |•, differs from distance $d[4, 1, 0]$, represented by • • |•, by exactly one block | • ••. Since this is true for all higher level distances, we can define a block value $\beta := \langle 4, c, h \rangle$ and use β to construct all distances up to level $4n - 2$, where the number of blocks varies between 1 and $n - 3$. In our example, this yields $d[8, 2, 1] = d[4, 1, 0] + \beta$. The multiplicities of the higher order distances can be determined straightforwardly. Finally, we define the (unique) distance of level $4n - 1$ explicitly as $d[4n - 1, n - 1, n] = \langle 4n - 1, (n - 1)c, nh \rangle + \varepsilon$. The resulting distances are shown in Figs. 9 and 10.

Our distance multiset D consists of all atoms z_i and c_i, and all distances as specified in the previous paragraphs, with the corresponding multiplicities. There are $4n - 1$ levels, and for each level ℓ 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.

Observe that the construction of D is possible for *any* instance of 3-Partition, and does *not* rely on the fact that there is a solution for the 3-Partition instance, nor on a particular ordering of the q_i's. In our argumentation above, we used

level ℓ	pseudo- atoms	multi- plicity	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	$12H$	$12H$	$d[3,0,1]$	$\langle 3,0,12H\rangle+\varepsilon$
	|●●	$n-1$	$6H$	$8H$	$d[3,1,0]$	$\langle 3,c,9H\rangle$
	●|●	$n-1$	$7H$	$10H$	$d[3,1,0]$	
	●●|	$n-1$	$8H$	$12H$	$d[3,1,0]$	
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	●●●|●●●	$n-1$	$24H$	$24H$	$d[7,1,2]$	$\langle 7,c,24H\rangle$
	●●|●●●|	$n-2$	$20H$	$24H$	$d[7,2,1]$	$\langle 7,2c,21H\rangle$
	●|●●●|●	$n-2$	$19H$	$22H$	$d[7,2,1]$	
	|●●●|●●	$n-2$	$18H$ 32	$20H$	$d[7,2,1]$	

Fig. 8. Distances up to level 7.

these two properties of the instance only to construct simultaneously a point set P that matches D up to error ε. Hence, we have constructed an instance D and ε of Partial Digest with Additive Errors from the given instance of 3-Partition, and we have shown already that a solution for the 3-Partition instance yields a solution for the Partial Digest with Additive Errors instance.

In the following, we show the opposite direction, i.e., we show that a solution for the Partial Digest with Additive Errors 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 Partial Digest with Additive Errors instance, i.e., multiset D is the multiset of pairwise distances of R, up to error ε for each distance. We assume w.l.o.g. that the points are ordered from left to right, i.e., $r_1<r_2<\cdots<r_{4n}$. We will show that R is basically identical to P, the point set that we constructed above.

level	pseudo-	multi-	distance	distance
ℓ	atoms	plicity	name	value
$4k+4$	$\bullet\bullet\vert\ldots\vert\bullet$	$n-k-1$	$d\,[4+4k,1+k,0+k]$	$d\,[4,1,0]+k\cdot\beta$
	$\bullet\vert\ldots\vert\bullet\bullet$	$n-k-1$	$d\,[4+4k,1+k,0+k]$	
	$\bullet\bullet\bullet\vert\ldots\vert$	$n-k-1$	$d\,[4+4k,1+k,1+k]$	$d\,[4,1,1]+k\cdot\beta$
	$\vert\ldots\vert\bullet\bullet\bullet$	$n-k-1$	$d\,[4+4k,1+k,1+k]$	
$5+4k$	$\bullet\bullet\vert\ldots\vert\bullet\bullet$	$n-k-1$	$d\,[5+4k,1+k,0+k]$	$d\,[5,1,0]+k\cdot\beta$
	$\bullet\bullet\bullet\vert\ldots\vert\bullet$	$n-k-1$	$d\,[5+4k,1+k,1+k]$	$d\,[5,1,1]+k\cdot\beta$
	$\bullet\vert\ldots\vert\bullet\bullet\bullet$	$n-k-1$	$d\,[5+4k,1+k,1+k]$	
	$\vert\ldots\vert\bullet\bullet\bullet\vert$	$n-k-2$	$d\,[5+4k,2+k,1+k]$	$d\,[5,2,1]+k\cdot\beta$
$6+4k$	$\bullet\bullet\bullet\vert\ldots\vert\bullet\bullet$	$n-k-1$	$d\,[6+4k,1+k,1+k]$	$d\,[6,1,1]+k\cdot\beta$
	$\bullet\bullet\vert\ldots\vert\bullet\bullet\bullet$	$n-k-1$	$d\,[6+4k,1+k,1+k]$	
	$\bullet\vert\ldots\vert\bullet\bullet\bullet\vert$	$n-k-2$	$d\,[6+4k,2+k,1+k]$	$d\,[6,2,1]+k\cdot\beta$
	$\vert\ldots\vert\bullet\bullet\bullet\vert\bullet$	$n-k-2$	$d\,[6+4k,2+k,1+k]$	
$7+4k$	$\bullet\bullet\bullet\vert\ldots\vert\bullet\bullet\bullet$	$n-k-1$	$d\,[7+4k,1+k,2+k]$	$d\,[7,1,2]+k\cdot\beta$
	$\bullet\bullet\vert\ldots\vert\bullet\bullet\bullet\vert$	$n-k-2$	$d\,[7+4k,2+k,1+k]$	$d\,[7,2,1]+k\cdot\beta$
	$\bullet\vert\ldots\vert\bullet\bullet\bullet\vert\bullet$	$n-k-2$	$d\,[7+4k,2+k,1+k]$	
	$\vert\ldots\vert\bullet\bullet\bullet\vert\bullet\bullet$	$n-k-2$	$d\,[7+4k,2+k,1+k]$	

Fig. 9. Distances with level 8 to $4n-5$. Value k varies between 1 and $n-3$.

level	lower	upper	distance	distance
ℓ	bound	bound	name	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	$d\,[4n-1,n-1,n]$	$\langle 4n-1,(n-1)c,nh\rangle+\varepsilon$

Fig. 10. Distances with level $4n-4$ to $4n-1$. Each case occurs once.

Obviously, error ε 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 $+\varepsilon$ 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

$$\begin{aligned}\sum_{i=1}^{3n} z_i + \sum_{i=1}^{n-1} c_i &= \sum_{i=1}^{3n} (\langle 1, 0, q_i\rangle - \varepsilon) + \sum_{i=1}^{n-1} (\langle 1, c, 0\rangle - \varepsilon)\\ &= \langle 3n, 0, nh\rangle - 3n\varepsilon + \langle n-1, (n-1)c, 0\rangle - (n-1)\varepsilon\\ &= \langle 4n-1, (n-1)c, nh\rangle - (4n-1)\varepsilon.\end{aligned}$$

On the other hand, the largest distance in multiset D is $d[4n-1, n-1, n] = \langle 4n-1, (n-1)c, nh\rangle + \varepsilon$. Each atom is the distance between two adjacent points in R, up to error ε, while $d[4n-1, n-1, n]$ is the distance between the first and the last point in R, again up to error ε. Hence, the atoms must sum up to the length of the largest distance. This is only possible if we apply error $+\varepsilon$ to each atom, yielding sum $\langle 4n-1, (n-1)c, nh\rangle$, and if we apply error $-\varepsilon$ to the largest distance, yielding $\langle 4n-1, (n-1)c, nh\rangle$ as well. Knowing this, we can again define *pseudoatoms* $\hat{z}_i = z_i + \varepsilon$ and $\hat{c}_i = c_i + \varepsilon$, 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 $1 \leqslant i \leqslant 3n$.

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 $\hat{c}_i$. Again, we call the pseudoatoms with value c in the second digit c-pseudoatoms, and those with value 0 in the second digit are called z-pseudoatoms. 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 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 $nh + n\varepsilon$. 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 $-\varepsilon$. This implies that each three q_i's that correspond to one block sum up to exactly h (since we have applied error $+\varepsilon$ to each atom to define the z-pseudoatoms). Thus, these triples yield a solution for the 3-Partition instance. □

5. Strong NP-completeness of Partial Digest with Relative Error

In this section, we show that Partial Digest with Relative Error is strongly NP-complete by using a reduction from 3-Partition similar to the one used to prove strong NP-completeness of Partial Digest with Additive Errors (see Theorem 4.1).

Theorem 5.1. Partial Digest with Relative Error *is strongly* NP-*complete, even if the error is a constant.*

Proof. The problem is in NP analogously to the proof of Theorem 4.1. The proof of NP-hardness is also along the lines of the proof of Theorem 4.1. 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 forms an instance of Partial Digest with Relative Error. The distances are expressed as numbers with a base Z, with $Z = 10hnc$ and $c = n^2h^2$.

In Partial Digest with Relative Error, all distances are required to be integers. For the purposes of this proof, we first relax this condition and show how the proof works if we allow even real numbers in the input of Partial Digest with Relative Error. In a second step, we will show how we can guarantee that all distances are integers. [5]

[5] Note that the choice of values Z and c is not tight, i.e., even smaller values might be possible.

We replace the definition of the atoms as follows:

$$z_i = \langle 1, 0, q_i\rangle \cdot \frac{1}{1+\mathrm{e}} \quad \text{for } 1 \leqslant i \leqslant 3n,$$
$$c_i = \langle 1, c, 0\rangle \cdot \frac{1}{1+\mathrm{e}} \quad \text{for } 1 \leqslant i \leqslant n-1,$$

where $\mathrm{e} = \frac{1}{100}$.

All z_i's and c_i's are part of the distance set E. Note that for a fixed level ℓ, the corresponding distances $d[\ell, \cdot, \cdot]$ from the proof of Theorem 4.1 are defined for at most two consecutive values of the second digit, say $d(\ell, j, \cdot)$ and $d(\ell, j+1, \cdot)$. For the relative error proof, we define distances $e[\ell, j]$ and $e[\ell, j+1]$ for all levels $2 \leqslant \ell \leqslant 4n-1$ and corresponding j or $j+1$, respectively, as follows. Here, we use values $B_u()$ and $B_l()$ which are specified below

$$e[\ell, j] = \langle \ell, j, B_u(\ell, j)\rangle \cdot \frac{1}{1+\mathrm{e}},$$
$$e[\ell, j+1] = \langle \ell, j+1, B_l(\ell, j+1)\rangle \cdot \frac{1}{1-\mathrm{e}}.$$

The first digit ℓ 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 Figs. 8–10, taken over all distances $d[\ell, j, \cdot]$ (for Fig. 9, these bounds result from Fig. 8 by adding appropriate multiples of h); similarly, value $B_l(\ell, j+1)$ is the minimum lower bound from the corresponding column in the same 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 $e[5, 1] = \langle 5, 1, 20H\rangle \cdot \frac{1}{1+\mathrm{e}}$ with multiplicity $3(n-1)$, while $e[6, 2] = \langle 6, 2, 15H\rangle \cdot \frac{1}{1-\mathrm{e}}$ 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 $e[4\ell', j] = \langle 4\ell', j, B_u(4\ell', j)\rangle \cdot 1/1 - \mathrm{e}$.

Finally, we define two special distances:

$$e[3, 0] = \langle 3, 0, h\rangle \cdot \frac{1}{1+\mathrm{e}},$$
$$e[4n-1, n-1] = \langle 4n-1, (n-1)c, nh\rangle \cdot \frac{1}{1-\mathrm{e}}.$$

Here, $e[3, 0]$ has multiplicity n, and distance $e[4n-1, n-1]$ has multiplicity 1.

All the distances, including the atoms, are put into distance multiset E. This completes our description of how to construct a Partial Digest with Relative Error instance from a given 3-Partition instance. We now show that a solution for the 3-Partition instance yields a solution for the Partial Digest with Relative Error instance, and vice versa; thereafter, we will present a strategy how to transform these rational distances into integer distances.

We first show that a feasible solution for the 3-Partition instance can be turned into a feasible solution for the Partial Digest with Relative Error instance. Assume that we are given a solution for the 3-Partition instance. Assume w.l.o.g. that the 3-Partition solution is such that the correct triples are formed by the numbers (q_1, q_2, q_3), $(q_4, q_5, q_6), \ldots, (q_{3n-2}, q_{3n-1}, q_{3n})$, and $q_i \leqslant q_{i+1} \leqslant q_{i+2}$ within each triple. We now define a set $P = \{p_1, \ldots, p_{4n}\}$ of points on a line, and show that P is a solution for the Partial Digest with Relative Error instance. Let $p_{u,v}$ denote the distance between points p_u and p_v. In order to be able to distinguish the distances from multiset E from distances $p_{u,v}$, we refer to distances $p_{u,v}$ as *point distances*.

We define all point distances by setting the atomic point distances $p_{u,u+1}$ as follows for integer u' with $0 \leqslant u' \leqslant n-1$:

$$p_{4u',4u'+1} = z_{3u'+1} \cdot (1+\mathrm{e}),$$
$$p_{4u'+1,4u'+2} = z_{3u'+2} \cdot (1+\mathrm{e}),$$
$$p_{4u'+2,4u'+3} = z_{3u'+3} \cdot (1+\mathrm{e})$$

while, for $0 \leqslant u' \leqslant n-2$, we let

$$p_{4u'+3,4u'+4} = c_{u'} \cdot (1+\mathrm{e}).$$

In other words, we apply the maximum error $1+\mathrm{e}$ to all z- and c-atoms and order them as $z_1, z_2, z_3, c_1, z_4, z_5, z_6, c_2, \ldots, c_{n-1}, z_{3n-2}, z_{3n-1}, z_{3n}$. Observe that this is the same ordering as in the proof of Theorem 4.1.

In order to show that this is a feasible solution for the Partial Digest with Relative Error instance, we have to indicate how the distances from multiset E are matched within the allowable error range to point distances $p_{u,v}$. As indicated by the definition of point set P, the z-atoms and c-atoms from E are matched to atomic point distances $p_{u,u+1}$, applying error $1+\mathrm{e}$. The maximum distance $e[4n-1, n-1]$ is matched to $p_{1,4n}$ with error $1-\mathrm{e}$. Any other distance $e[\ell, j]$ (resp. $e[\ell, j+1]$) is matched to any point distance $p_{u,u+\ell}$ spanning exactly ℓ atoms, of which j (resp. $j+1$) are c-atoms, using an appropriate relative error. In order to see that this is feasible, we need to show that error e is large enough such that distance $e[\ell, j]$ can be used to generate every possible distance spanning over ℓ atoms.

More precisely, if distance $e[\ell, j]$ is matched to point distance $p_{u,u+\ell}$, it suffices to show that $e[\ell, j](1-\mathrm{e}) \leqslant p_{u,u+\ell} \leqslant e[\ell, j](1+\mathrm{e})$. By construction of the matching, the number of c-atoms between p_u and $p_{u+\ell}$ is j. Hence, we have $p_{u,u+\ell} = \langle \ell, jc, S\rangle$, where S denotes the sum of the q_i's corresponding to the z-atoms between p_u and $p_{u+\ell}$. Since $S \leqslant B_u(\ell, j)$ by construction, we have immediately $p_{u,u+\ell} \leqslant e[\ell, j](1+\mathrm{e})$. On the other hand, with $\mathrm{e} = \frac{1}{100} \geqslant \frac{nh}{2Z^2} \Rightarrow (\ell Z^2 + jcZ + nh)(1-\mathrm{e}) \leqslant (\ell Z^2 + jcZ)(1+\mathrm{e})$, we have $e[\ell, j](1-\mathrm{e}) \leqslant p_{u,u+\ell}$.

Analogously, it can be shown that error e is sufficient to match distance $e[\ell, j+1]$ to point distance $p_{u,u+\ell}$, where we use that the sum of the corresponding q_i's is at least $B_l(\ell, j+1)$. For the special cases $e[3,0]$, the matching point distance is exactly $\langle 3, 0, h\rangle$, since the three q_i corresponding to the z-atoms belong to the same triple and sum up to h.

We now show the opposite direction of our claim, i.e., we show that a feasible solution for the Partial Digest with Relative Error instance can be turned into a feasible solution for the 3-Partition instance. We assume that we are given a solution of the Partial Digest with Relative Error instance as a set of points $P = \{p_1, \ldots, p_{4n}\}$ on a line. Let $p_{u,v}$ again denote the point distance between points p_u and p_v. As the solution is feasible, each point distance $p_{u,v}$ is matched to a distance $e[\ell, \cdot]$ or to an atom z_i or c_i (in a bijective way).

The following five item show how to construct a 3-Partition solution:

(1) The atomic point distances $p_{u,u+1}$ for $0 \leqslant u \leqslant 4n-2$ must be matched to the atoms z_i and c_i in any solution and any matching.

Proof. We first show that no non-atomic distance $e[\ell, \cdot]$ from multiset E can be made smaller by using the error range than any z- or c-atom, even if we make the atom as large as possible. To this end, we show that $z_i \cdot (1+\mathrm{e}) < e[\ell, \cdot] \cdot (1-\mathrm{e})$ for any z-atom z_i, and $c_i \cdot (1+\mathrm{e}) < e[\ell, \cdot] \cdot (1-\mathrm{e})$ for any c-atom c_i: we have $z_i = (Z^2 + q_i) \cdot \frac{1}{1+\mathrm{e}}$ and $c_i = (Z^2 + Z) \cdot \frac{1}{1+\mathrm{e}}$; with $q_i \leqslant \frac{Z^2}{2}$ and $Z \leqslant \frac{Z^2}{2}$ (by definition of Z), we have that z_i and c_i are both smaller or equal to $\frac{3}{2}Z^2 \cdot \frac{1}{1+\mathrm{e}}$. Thus, it is sufficient to show that $\frac{3}{2}Z^2 \cdot \frac{1}{1+\mathrm{e}} < \ell Z^2 \cdot \frac{1-\mathrm{e}}{1+\mathrm{e}}$, where the right side models a lower bound for $e[\ell, \cdot]$. Straight-forward analysis shows that this is true if $\mathrm{e} < \frac{1}{4}$, which is true by definition of e.

Let us now assume for the sake of contradiction that point distance $p_{u',u'+1}$ is matched to a non-atomic distance $e[\ell, j]$. This implies that in the matching at least one atom is matched to a non-atomic point distance $p_{u,v}$ with $v \geqslant u+2$. The atomic point distances $p_{u,u+1}$ and $p_{u+1,u+2}$ must be matched with some distances from E; since the z-atoms are the smallest distances among all distances in E, we have that both $p_{u,u+1}$ and $p_{u+1,u+2}$ are at least $Z^2 \cdot \frac{1-\mathrm{e}}{1+\mathrm{e}}$ (applying error $1-\mathrm{e}$), hence, $p_{u,v} \geqslant 2Z^2 \cdot \frac{1-\mathrm{e}}{1+\mathrm{e}}$. On the other hand, point distance $p_{u,v}$ is matched to some atom a, hence $p_{u,v} \leqslant a \cdot (1+\mathrm{e})$. Since each atom is smaller than $(Z^2 + cZ + \frac{h}{2}) \cdot \frac{1}{1+\mathrm{e}}$, we have $2Z^2 \cdot \frac{1-\mathrm{e}}{1+\mathrm{e}} \leqslant p_{u,v} \leqslant a \cdot (1+\mathrm{e}) < (Z^2 + cZ + \frac{h}{2}) \cdot \frac{1+\mathrm{e}}{1+\mathrm{e}}$. With $c < \frac{Z}{4}$ and $\frac{h}{2} < \frac{Z^2}{4}$ (by definition), this yields $\mathrm{e} > \frac{1}{7}$, which contradicts our definition of e, which was set to $\frac{1}{100}$. □

(2) Any solution has to use the maximum error $1+\mathrm{e}$ for all atoms z_i and c_i, i.e., $p_{u,u+1} = z_i \cdot (1+\mathrm{e})$ or $p_{u,u+1} = c_i \cdot (1+\mathrm{e})$.

Proof. To see this, we first show that the sum of all atoms z_i and c_i multiplied by $1+\mathrm{e}$ is equal to the longest distance $e[4n-1, n-1]$ multiplied by $1-\mathrm{e}$: $(\sum_{i=1}^{3n} z_i + \sum_{i=1}^{n-1} c_i) \cdot (1+\mathrm{e}) = \sum_{i=1}^{3n} \langle 1, 0, q_i\rangle + \sum_{i=1}^{n-1} \langle 1, c, 0\rangle = \langle 4n-1, (n-1)c, nh\rangle = e[4n-1, n-1] \cdot (1-\mathrm{e})$. Together with the fact that the c_i's and z_i's are matched to atomic point distances (due to Item 1), this implies that in any solution and any matching, error $1+\mathrm{e}$ has to be applied to each atom (and error $1-e$ to the maximum distance) in order to guarantee that the maximum distance can be matched to some point distance. □

(3) In any solution, any distance $e[\ell, \cdot]$ must match a point distance $p_{u,u+\ell}$ for $\ell < \frac{1}{4\mathrm{e}} - 1$ (i.e., parameter ℓ actually describes the level of the distance in any solution and any matching, for small values of ℓ).

Proof. Let $\ell < \frac{1}{4\mathrm{e}} - 1$. We first show that no distance $e[\ell', \cdot]$ with $\ell' < \ell$ can match a point distance $p_{u,u+\ell}$, even if maximum error $1 + \mathrm{e}$ is applied. To see this, we prove that $e[\ell', \cdot] \cdot (1 + \mathrm{e}) < p_{u,u+\ell}$: first, observe that $e[\ell', \cdot] \leqslant \langle \ell', (n-1)c, nh\rangle \cdot \frac{1}{1-\mathrm{e}}$. Thus, with $(n-1)c < \frac{Z}{4}$ and $nh < \frac{Z^2}{4}$ (by definition of c and Z), we have $e[\ell', \cdot] \cdot (1+\mathrm{e}) \leqslant \langle \ell', (n-1)c, nh\rangle \cdot \frac{1+\mathrm{e}}{1-\mathrm{e}} = (\ell' Z^2 + (n-1)cZ + hn) \cdot \frac{1+\mathrm{e}}{1-\mathrm{e}} \leqslant (\ell' + \frac{1}{2})Z^2 \cdot \frac{1+\mathrm{e}}{1-\mathrm{e}}$. Since $\ell' < \ell$, we have that $(\ell' + \frac{1}{2}) \cdot \frac{1+\mathrm{e}}{1-\mathrm{e}} < \ell' + 1$ implies $e[\ell', \cdot] \cdot (1+\mathrm{e}) < (\ell'+1)Z^2 \leqslant \ell Z^2$. On the other hand, we have $p_{u,u+\ell} \geqslant \ell Z^2$, since each atomic point distance within $p_{u,u+\ell}$ has at least size Z^2 (due to Item 1).

We now show that no distance $e[\ell', \cdot]$ with $\ell' > \ell$ can match a point distance $p_{u,u+\ell}$, even if minimum error $1 - \mathrm{e}$ is applied. To this end, we prove that $e[\ell', \cdot] \cdot (1 - e) > p_{u,u+\ell}$: first, we have $e[\ell', \cdot] \cdot (1 - \mathrm{e}) \geqslant \ell' Z^2 \cdot \frac{1-\mathrm{e}}{1+\mathrm{e}}$ by definition of distance $e[\ell, \cdot]$. Second, we have $p_{u,u+\ell} \leqslant \ell Z^2 + (n-1)cZ + nh < (\ell + \frac{1}{2})Z^2$, since each atomic point distance has at most size $\ell Z^2 + (n-1)cZ + nh$. Finally, since $\ell < \frac{1}{4\mathrm{e}} - 1$, we have $(\ell + \frac{1}{2}) \cdot \frac{1+\mathrm{e}}{1-\mathrm{e}} < \ell + 1$. Combining these three inequalities yields that $e[\ell', \cdot] \cdot (1 - e) > p_{u,u+\ell}$.

The previous two paragraphs prove that no distance $e[\ell', \cdot]$ with $\ell' \neq \ell$ can match point distance $p_{u,u+\ell}$ in any solution and any matching. □

(4) Any solution and corresponding matching must be such that atomic point distances

$$p_{4u',4u'+1}, \quad p_{4u'+1,4u'+2}, \quad p_{4u'+2,4u'+3}$$

for $0 \leqslant u' \leqslant n-1$ are matched to z-atoms, and atomic point distances $p_{4u'+3,4u'+4}$ for $0 \leqslant u' \leqslant n-2$ are matched to c-atoms. In other words, the ordering must of the form $zzz\,c\,zzz\,c\ \ldots\ c\,zzz$.

Proof. It suffices to show that any four adjacent atomic point distances $p_{u,u+1}, p_{u+1,u+2}, p_{u+2,u+3}, p_{u+3,u+4}$ cannot be all matched to z-atoms, because only $4n - 1$ atoms exist, and if two c-atoms were matched such that less than three z-atoms are inbetween, there would have to exist a location where at least four z-atoms occurred in a row, in contradiction to the claim.

Assume for the sake of contradiction that atomic point distances $p_{u,u+1}, p_{u+1,u+2}, p_{u+2,u+3}, p_{u+3,u+4}$ are all matched to z-atoms, say w.l.o.g. z_1, z_2, z_3, z_4. We consider point distance $p_{u,u+4}$; according to Item 3, this point distance must be matched to a distance $e[4, 1]$, which is the only distance of type $e[4, \cdot]$ that exists in multiset E. It is a necessary condition that if the minimum error is applied to distance $e[4, \cdot]$, it must be smaller (or equal) to the sum of the four z-atoms multiplied by the maximum error; thus, using $q_i \leqslant 6H$, we have

$$\begin{aligned} e[4,1] \cdot (1-\mathrm{e}) &\leqslant (z_1 + z_2 + z_3 + z_4)(1+\mathrm{e}), \\ (4Z^2 + Z + 16H)\frac{1-\mathrm{e}}{1-\mathrm{e}} &\leqslant (4Z^2 + q_1 + q_2 + q_3 + q_4)\frac{1+\mathrm{e}}{1+\mathrm{e}}, \\ Z + 16H &\leqslant q_1 + q_2 + q_3 + q_4 \leqslant 24H, \\ Z &\leqslant 8H. \end{aligned}$$

The last inequality is false by definition (recall that $H = \frac{h}{12}$), thus we have shown a contradiction. □

(5) In any feasible solution three consecutive atomic point distances that are matched to z-atoms sum up to the desired value of h in the last digit.

Proof. Let p_u be any point where three consecutive atomic point distances start that are matched to z-atoms. Then $p_{u,u+3} = \langle 3, 0, S\rangle$, with S the sum of the three q_i's corresponding to the matched z-atoms, since error $1 + \mathrm{e}$ is applied in each z-atom, due to Item 1. Moreover, only distances with level $\ell = 3$ might match to point distance $p_{u,u+3}$, due to Item 3. Since $e[3, 1] \cdot (1 - \mathrm{e}) = \langle 3, c, 0\rangle > p_{u,u+3}$, only a distance $e[3, 0]$ can match point distance $p_{u,u+3}$. This implies that $\langle 3, 0, S\rangle = p_{u,u+3} \leqslant e[3, 0](1 + \mathrm{e}) = \langle 3, 0, h\rangle$, hence $S \leqslant h$. Since the previous inequality holds for *any* triple of adjacent atomic point distances that are matched to z-atoms, and there are n such triples due to Item 4, summing up all these triples yields at most total nh in the third digit. On the other hand, since the q_i's sum up to $\sum_{i=1} 3n\, q_i = nh$

as well (by definition of 3-Partition), the previous inequalities need to be tight, hence, each consecutive atomic point distances that are matched to z-atoms sum up to value h in the last digit. □

To finish our proof, we show how to make all distances integer. First, observe that all distances in E are integers multiplied by either $\frac{1}{1-\mathrm{e}}$ or $\frac{1}{1+\mathrm{e}}$. Since error e is a rational number, we can express it as $\mathrm{e} = \frac{p}{q}$, with p and q co–primes (i.e., the greatest common divisor of p and q is 1). Then $\frac{1}{1+\mathrm{e}} = \frac{q}{p+q}$ and $\frac{1}{1-\mathrm{e}} = \frac{q}{q-p}$. Hence, if we multiply *all* distances by $(p+q)(q-p) = q^2 - p^2$, then we obtain integer distances, and they still fulfill all properties above, since each distance is scaled by the same factor.

The distance multiset E contains $4n-1$ distances, and each distance is a 3-digit number with base Z. By construction, the largest distance is $d[4n-1, n-1, n] = \langle 4n-1, (n-1)c, nh\rangle \cdot \frac{1}{1-\mathrm{e}} \cdot (q^2 - p^2)$. With $\mathrm{e} = \frac{p}{q}$, we have $d[4n-1, n-1, n] = \langle 4n-1, (n-1)c, nh\rangle \cdot q(p+q)$. Since $\mathrm{e} = \frac{1}{100}$, $Z = 10hnc$, and $c = n^2h^2$, each distance in E is polynomially bounded in n and h. Hence, we can construct multiset E and error e from a given instance of 3-Partition in polynomial time, and our reduction shows that Partial Digest with Relative Error is strongly NP-hard. □

6. Conclusion

We have shown that the minimization problem Min Partial Digest Superset is NP-hard, and that the maximization problem Max Partial Digest Subset is hard to approximate. This partially answers open problem 12.116 left open in the book by Pevzner [24], as our results rule out the possibility of having exact polynomial-time algorithms. Moreover, we have shown that Partial Digest is strongly NP-complete if all measurements are prone to the same additive or multiplicative error. However, in the realm of Partial Digest, many questions are still open:

- Since our optimization variations model different error types that (always) occur in real-life data, our hardness results suggest that real–life Partial Digest problems are in fact instances of NP-hard problems. However, the backtracking algorithm from [19] performs well in experiments [35]. How can this be explained?
- What is the best approximation ratio for Min Partial Digest Superset?
- In our NP-hardness proof for Partial Digest with Additive Errors, we used non-constant error $\varepsilon = \frac{h}{4}$. Is Partial Digest still NP-complete if we restrict the error to some (small) constant? What if we allow only one-sided errors, i.e., if the lengths of the distances are for instance always underestimated?
- Using gel electrophoresis, it is very hard to determine the correct multiplicity of a distance. This yields the following variation of Partial Digest: we are given a *set* of distances, and for each distance a multiplicity, and we ask for points on a line such that the multiplicities of the corresponding distance set do not differ "too much" from the given multiplicities. What is the computational complexity of this problem?
- Is there a polynomial-time algorithm for the Partial Digest problem if we restrict the input to be a *set* of distances (instead of a multiset), i.e., if we know in advance that each two distances in the input are pairwise distinct?

Finally and obviously, the main open problem is of course the computational complexity of Partial Digest itself.

Acknowledgments

Preliminary versions of parts of this work have appeared as a technical report [8] and in the conference proceedings of WABI 2003 [9] and LATIN 2004 [7]. A part of this work was done while S. Eidenbenz and P. Penna were working as postdocs at ETH Zurich and a part of the work was done while M. Cieliebak was visiting Los Alamos National Laboratory. We would like to thank Claudio Gutiérrez, Fabian Hennecke, Roland Ulber, Birgitta Weber, and Peter Widmayer for helpful discussions, Riko Jacob, who suggested the graphical presentation used in Section 4, and anonymous referees whose comments helped improve the presentation of this paper.

References

[1] F. Alizadeh, R.M. Karp, L.A. Newberg, D.K. Weisser, Physical mapping of chromosomes: a combinatorial problem in molecular biology, in: Proc. of the Fourth SIAM–ACM Symp. on Discrete Algorithms (SODA 1993), 1993, pp. 371–381.

[2] L. Allison, C.N. Yee, Restriction site mapping is in separation theory, Comput. Appl. Biosci. 4 (1) (1988) 97–101.

[3] V. Bafna, N. Edwards, On de novo interpretation of tandem mass spectra for peptide identification, in: Proc. of the Seventh Annu. Internat. Conf. on Computational Biology (RECOMB 03), 2003, pp. 9–18.

[4] S. Baginsky, Personal Communication, ETH Zurich, Institute of Plant Sciences, 2003.

[5] J. Błażewicz, P. Formanowicz, M. Kasprzak, M. Jaroszewski, W.T. Markiewicz, Construction of DNA restriction maps based on a simplified experiment, Bioinformatics 17 (5) (2001) 398–404.

[6] T. Chen, M.-Y. Kao, M. Tepel, J. Rush, G.M. Church, A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry, in: Proc. of the 11th SIAM–ACM Symp. on Discrete Algorithms (SODA 2000), 2000, pp. 389–398.

[7] J. Błażewicz, P. Formanowicz, M. Kasprzak, M. Jaroszewski, W.T. Markiewicz, Construction of DNA restriction maps based on a simplified experiment, Bioinformatics 17 (5) (2001) 398–404.

[8] M. Cieliebak, S. Eidenbenz, P. Penna, Noisy data make the partial digest problem NP-hard, Technical Report 381, ETH Zurich, Department of Computer Science, 2002.

[9] M. Cieliebak, S. Eidenbenz, P. Penna, Noisy data make the partial digest problem NP-hard, in: Proc. of the Third Workshop on Algorithms in Bioinformatics (WABI 2003), 2003, pp. 111–123.

[10] T. Dakić, On the turnpike problem, Ph.D. Thesis, Simon Fraser University, 2000.

[11] T.I. Dix, D.H. Kieronska, Errors between sites in restriction site mapping, Comput. Appl. Biosci. 4 (1) (1988) 117–123.

[12] D. Fasulo, Algorithms for DNA restriction mapping, Ph.D. Thesis, University of Washington, 2000.

[13] J. Fütterer, Personal Communication, ETH Zurich, Institute of Plant Sciences, 2002.

[14] M.R. Garey, D.S. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness, Freeman, New York, 1979.

[15] L. Goldstein, M.S. Waterman, Mapping DNA by stochastic relaxation, Adv. in Appl. Math. 8 (1987) 194–207.

[16] J. Håstad, Clique is hard to approximate within $n^{1-\varepsilon}$, Acta Mathematica 182 (1999) 105–142.

[17] J. Inglehart, P.C. Nelson, On the limitations of automated restriction mapping, Comput. Appl. Biosci. 10 (3) (1994) 249–261.

[18] P. James (Ed.), Proteome Research: Mass Spectrometry, Springer, Berlin, 2001.

[19] P. Lemke, S.S. Skiena, W. Smith, Reconstructing sets from interpoint distances, Technical Report TR2002-37, DIMACS, 2002.

[20] P. Lemke, 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.

[21] E.W. Mayr, H.J. Prömel, A. Steger, Lectures on Proof Verification and Approximation Algorithms, Springer, Berlin, 1998.

[22] L. Newberg, D. Naor, A lower bound on the number of solutions to the probed partial digest problem, Adv. in Appl. Math. 14 (1993) 172–183.

[23] G. Pandurangan, H. Ramesh, The restriction mapping problem revisited, J. Comput. System Sci. 65 (3) (2002) 526–544 special issue on Computational Biology.

[24] P.A. Pevzner, Computational Molecular Biology: An Algorithmic Approach, MIT Press, Cambridge, MA, 2000.

[25] P.A. Pevzner, M.S. Waterman, Open combinatorial problems in computational molecular biology, in: Proc. of the Third Israel Symp. on Theory of Computing and Systems (ISTCS 1995), 1995, pp. 158–173.

[26] J. Rosenblatt, P. Seymour, The structure of homometric sets, SIAM J. Algorithms Discrete Math. 3 (3) (1982) 343–350.

[27] D.B. Searls, Formal grammars for intermolecular structure, in: Proc. of the First Internat. Symp. on Intelligence in Neural and Biological Systems (INBS'95), 1995, pp. 30–37.

[28] J. Setubal, J. Meidanis, Introduction to Computational Molecular Biology, PWS, Boston, MA, 1997.

[29] S.S. Skiena, W. Smith, P. Lemke, Reconstructing sets from interpoint distances, in: Proc. of the Sixth ACM Symp. on Computational Geometry (SoCG 1990), 1990, pp. 332–339.

[30] S.S. Skiena, G. Sundaram, A partial digest approach to restriction site mapping, Bull. Math. Biol. 56 (1994) 275–294.

[31] P. Tuffery, P. Dessen, C. Mugnier, S. Hazout, Restriction map construction using a 'complete sentence compatibility' algorithm, Comput. Appl. Biosci. 4 (1) (1988) 103–110.

[32] M.S. Waterman, Introduction to Computational Biology, Chapman & Hall, New York, 1995.

[33] G.J. Woeginger, Z.L. Yu, On the equal-subset-sum problem, Inform. Process. Lett. 42 (1992) 299–302.

[34] L.W. Wright, J.B. Lichter, J. Reinitz, M.A. Shifman, K.K. Kidd, P.L. Miller, Computer–assisted restriction mapping: an integrated approach to handling experimental uncertainty, Comput. Appl. Biosci. 10 (4) (1994) 435–442.

[35] Z. Zhang, An exponential example for a partial digest mapping algorithm, J. Comput. Biol. 1 (3) (1994) 235–239.