Noisy Data Make the Partial Digest Problem NP-hard

TECHNICAL REPORT no. 381
ETH Zurich, Department of Computer Science

Mark Cieliebak
cieliebak@inf.ethz.ch

Stephan Eidenbenz
eidenben@inf.ethz.ch

Paolo Penna
penna@inf.ethz.ch

Abstract

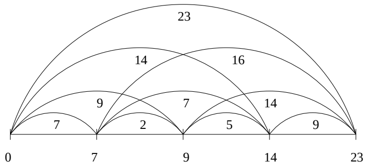
The Partial Digest problem – well-known for its applications in computational biology and for the intriguingly open status of its computational complexity – asks for the coordinates of n points on a line such that the pairwise distances of the points form a given multi-set of $\binom{n}{2}$ distances. In an effort to model real-life data, we study the computational complexity of a minimization version of Partial Digest, in which only a subset of all pairwise distances is given and the rest are lacking due to experimental errors. We show that this variation is NP-hard to solve exactly, thus making the existence of polynomial-time algorithms for this problem extremely unlikely. Our result answers an open question posed by Pevzner (2000). We then study a maximization version of Partial Digest where a superset of all pairwise distances is given, with some additional distances due to inaccurate measurements. We show that this maximization version is NP-hard to approximate to within a factor of $|D|^{\frac{1}{2}-\epsilon}$ for any $\epsilon > 0$, where $|D|$ is the number of input distances, which implies that polynomial-time algorithms cannot even guarantee to find a solution for the problem that comes close to the optimum. Our inapproximability result is tight up to low-order terms as we give a trivial approximation algorithm that achieves a matching approximation ratio. Our optimization variations model two different error types that occur in real-life data.

Keywords: Partial Digestion, Turnpike Problem, Computational Biology

1 Introduction

The Partial Digest problem is one of the most intriguing problems from computational biology: on the one hand, it is a basic problem with relevant applications in DNA sequencing; on the other hand, its computational complexity is a long–standing 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 such that D is exactly the multiset of all pairwise distances of these points.

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



More formally, the Partial Digest problem can be defined as follows.

Definition (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 $\{|p_i - p_j| \mid 1 \leq i < j \leq m\} = D$?*

This problem has – among others – applications in the study of the structure of DNA molecules. Indeed, given a large DNA molecule, 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 at occurrences of the pattern *GAATTC*. Under appropriate experimental conditions (e.g. by exposing the enzyme for different time periods or by using very small amounts of the enzyme), *all* fragments between each two restriction sites are created. This process is called *partial digestion*, in contrast to *full digestion*, where the enzyme is applied long enough to cleave at all restriction sites. The lengths of the fragments (i.e., their number of nucleotides) are then measured (e.g. by using gel electrophoresis). 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.

In reality, the partial digest experiment cannot be conducted under ideal conditions as outlined above and thus errors occur in the data [17, 6]. A first source of erronous data is the digestion phase of the experiment: it can happen that a particular restriction site does not get cut in combination with all other restriction sites, but only in combination with some restriction sites; thus, some distances will be missing in the data. On the other hand, an enzyme may erronously cut 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. A second source of errors is the addition of distances through the insertion of third-party particles during the experiment, such as DNA from the staff. A third source of errors is the measurement phase: using gel electrophoresis, measurement errors within a range of upto 5 percent are very common; moreover, distances that do not occur in a large number of copies cannot be read with this measuring technique as they do not leave large enough spots, thus leading to omission of certain distances; furthermore, small fragments can be lost since they run of the end of the gel. Finally, determining the proper multiplicity of a fragment is a non–trivial problem. Hence, three types of errors occur: measurement errors, where the length of a distance is erroneous; additions, where additional distances occur that do not correspond to any fragment; and omissions, where distances of fragments are missing. In this paper, we define two optimization variations of Partial Digest, where the first variation models addition errors and the second variation models omission errors. Each variation allows only for one type of error to occur, and we will prove hardness results for both variations, implying that no polynomial-time algorithm can guarantee to find optimum or even nearly optimum solutions. Intuitively, the problem of modeling "real-life" instances – in which all three error types can occur – is even harder.

The Min Partial Digest Superset problem models the situation of omissions, where we are given data in which some distances are missing, and we search for a set of points such that the number of omitted distances is minimum. It is formally defined as follows.

Definition (Min Partial Digest Superset). *Given a multiset of k positive integers $D = \{d_1, \ldots, d_k\}$, find the minimum m such that there is a set of m integers $P = \{p_1, \ldots, p_m\}$ with $D \subseteq \{|p_i - p_j| \mid 1 \leq i < j \leq m\}$.*

The Max Partial Digest Subset problem models the situation of additions, where we are given data in which some wrong distances were added and we search for a set of points such that the number of added distances is minimum. A formal definition is as follows.

Definition (Max Partial Digest Subset). *Given a multiset of k positive integers $D = \{d_1, \ldots, d_k\}$, find the maximum m such that there is a set of m integers $P = \{p_1, \ldots, p_m\}$ with $\{|p_i - p_j| \mid 1 \leq i < j \leq m\} \subseteq D$.*

Our two 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 problems above whose optimum is $\binom{m}{2}$; any NO instance of Partial Digest is an instance of Max Partial Digest Subset (resp., Min Partial Digest Superset) whose optimum is at most $\binom{m}{2} - 1$ (resp., at least $\binom{m}{2} + 1$).

However, we show that such algorithms cannot exist unless $\mathsf{P} = \mathsf{NP}$: We first show that computing the optimal solution for the Min Partial Digest Superset problem is NP-hard, by proposing a reduction from the NP-hard problem Equal Sum Subsets. This implies that we could use a polynomial-time algorithm that solves Min Partial Digest Superset to solve Equal Sum Subsets in polynomial time as well, which could then be used to design polynomial-time algorithms for all NP-complete problems. Our result provides an answer to the open problem in [11, Problem 12.116], which asks for an algorithm to reconstruct a set of points, given a subset of their pairwise distances. We then strengthen our hardness result by considering the t-Partial Digest Superset problem, where we restrict the cardinality of a solution to at most t, for some fixed parameter t:

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

Clearly, the NP-hardness of Min Partial Digest Superset implies the NP-hardness of t-Partial Digest Superset, for some t. We show that the above problem remains NP-hard for *any* fixed $t = \Omega(|D|^{1/2+\epsilon})$ and any $\epsilon > 0$. This result is tight in a sense, since any solution (even from the original Partial Digest) must have at least cardinality $t = \Omega(|D|^{1/2})$.

As for the Max Partial Digest Subset problem, we show that there is no polynomial–time algorithm for this problem that guarantees to achieve an approximation ratio[1] of $|D|^{\frac{1}{2}-\epsilon}$ for any $\epsilon > 0$, unless P = NP, by proposing a reduction from Maximum Clique. We also point to a trivial approximation algorithm that achieves a matching approximation ratio. Thus, our result is tight up to low-order terms. Our inapproximability result means that not only can we not expect a polynomial-time algorithm that finds the optimum solution, but we cannot even expect a polynomial-time algorithm that finds solutions that are a factor $|D|^{\frac{1}{2}-\epsilon}$ off the optimum. The problem Maximum Clique is very hard to approximate and our reduction is gap-preserving (as introduced in [2]), thus transferring the inapproximability of Maximum Clique to Max Partial Digest Subset. Our hardness results show that a polynomial-time algorithm for the original Partial Digest (if any) cannot be obtained by looking at the two natural optimization problems we considered here. If any such algorithm exists, then it must exploit some combinatorial properties of Partial Digest instances that do not hold for these optimization problems.

The exact computational complexity of Partial Digest is a long–standing open problem: it can be solved in pseudopolynomial[2] time [13, 8], and there exists a backtracking algorithm (for exact or erroneous data) which has expected running time polynomial in the number of distances [16, 17], but exponential worst case running time [19]. If the points are not on a line but 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-completeness is known [10, 3, 11, 4, 15, 12]. Recently, the Partial Digest problem has received increasing attention due to its application in computational biology. However, in its pure combinatorial formulation it has been studied for a long time: It appears already in the 1930's in the sphere of X-ray crystallography (acc. to [16]); the problem is very closely related to the theory of homometric sets[3] [16]; it can be formalized by cut grammars, which have one additional symbol δ, the *cut*, that is neither a non–terminal nor a terminal symbol [14]; and finally, 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]. In the biological setting, many experimental variations have been studied: Double digestion, where two different enzymes are used [15]; probed partial digestion, where probes (markers) are hybridized to partially digested DNA [9, 1]; simplified partial digest, where clones are cleaved in either one or in all restriction sites [3]; labeled partial digestion, where both ends of the DNA molecule are labeled before digestion [10]; and multiple complete digestion, where many different enzymes are used [5]. For a good survey on the Partial Digest problem, see [16]; and

[1] The approximation ratio of an algorithm $\mathcal{A}$ for any instance I is $\frac{OPT(I)}{\mathcal{A}(I)}$, where $\mathcal{A}(I)$ is the number of points in the solution of algorithm $\mathcal{A}$, and $OPT(I)$ is the number of points in an optimal solution.

[2] I.e., polynomial in the largest number of the input, but not necessarily polynomial in the bit length of the largest number.

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

for more recent discussions on the problem, see [15] and [11].

The paper is organized as follows: In Sect. 2 we present the hardness results of Min Partial Digest Superset. Sect. 3 deals with the t-Partial Digest Superset problem. In Sect. 4 we provide the (in-) approximability results on Max Partial Digest Subset. Finally, we conclude and present some open problems in Sect. 5.

2 NP-hardness of Min Partial Digest Superset

In this section we show that Min Partial Digest Superset is NP-hard by proposing a reduction from Equal Sum Subsets. We start with some notation.

A *multiset* with elements $1, 1, 3, 5, 5$, and 8 is denoted by $\{1, 1, 3, 5, 5, 8\}$. Subtracting an element from a multiset will remove it only once (if it is there), thus $\{1, 1, 3, 5, 5, 8\} - \{1, 4, 5, 5\} = \{1, 3, 8\}$. Given a set of integers $X = \{x_1, \ldots, x_n\}$, the *distance multiset* $\Delta(X)$ is defined as the multiset of all distances of X, i.e., $\Delta(X) := \{|x_i - x_j| \mid 1 \leq i < j \leq n\}$. We denote the sum of the elements of a set X of integers by $\mathrm{sum}(X)$, i.e., $\mathrm{sum}(X) := \sum_{x \in X} x$. We say that a set of points P *covers* distance multiset D if $D \subseteq \Delta(P)$.

Let $D = \{d_1, \ldots, d_k\}$. If m is the minimal number such that a set P of cardinality m with $D \subseteq \Delta(P)$ exists, then $m \leq k + 1$: We set $p_0 = 0, p_i = p_{i-1} + d_i$, for $1 \leq i \leq k$, and $P_{triv} = \{p_0, \ldots, p_k\}$, i.e., we simply put all distances from D in a chain "one after the other" (cf. Figure 1). 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 $k + 1$.

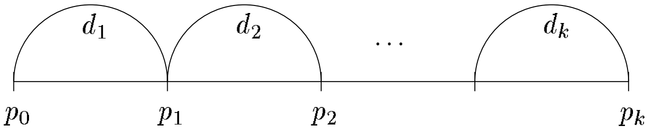


Figure 1: Trivial solution for a distance multiset D.

Observe that Partial Digest $\leq_p$ Min Partial Digest Superset: Given an instance P of Partial Digest of size $|P| = k$, there is a solution for P if and only if the minimal solution for the Min Partial Digest Superset instance P has size $m = \frac{1}{2} + \sqrt{\frac{1}{4} + 2k}$ (in this case, $k = \binom{m}{2}$).

Theorem 1. Min Partial Digest Superset *is* NP*-hard.*

Proof. We reduce Equal Sum Subsets to Min Partial Digest Superset, where Equal Sum Subsets is an NP-complete problem [18] that is defined as follows: Given a set of n numbers $A = \{a_1, \ldots, a_n\}$, are there two disjoint nonempty subsets $X, Y \subseteq A$ such that $\mathrm{sum}(X) = \mathrm{sum}(Y)$?

Given an instance $A = \{a_1, \ldots, a_n\}$ of Equal Sum Subsets, we set $D = A$ (and $k = n$), and claim the following: 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.

"only if" part: 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 $1 \leq r < s \leq n$.

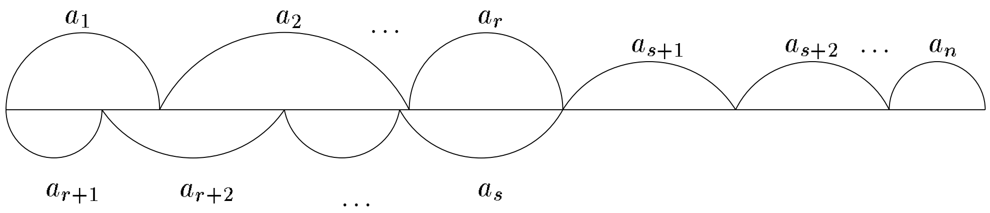


Figure 2: Solution if there are two sets of equal sum.

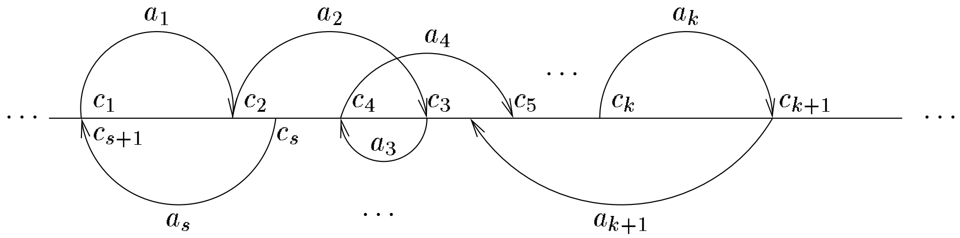


Figure 3: A solution containing a cycle yields two subsets of equal sum: the overall lenght of right jumps equals to the overall length of left jumps.

We construct a set P which covers D and 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 zero: those distances from X and those from Y (cf. Figure 2); the remaining distances from $D - (X \cup Y)$ are at the end of the two chains.

Notice that the two chains corresponding to X and Y share two points, namely zero and their common endpoint. This will allow us to find a covering set with at most n points. Let $P = \{p_0, \ldots, p_{s-1}, q_{s+1}, \ldots, q_n\}$. Obviously, P is a set of cardinality n. Moreover, by construction (cf. Figure 2), it holds that $D = \{a_1, \ldots, a_n\} \subseteq \Delta(P)$.

"if" part: 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 $m \leq n$, this graph contains a cycle. We show that such a cycle induces a solution of the Equal Sum Subsets instance.

Let $C = c_1, \ldots, c_s$ be a cycle in G (see Fig. 3). Then $|c_{i+1} - c_i| \in D$, for all $1 \leq i \leq s$ (with some abuse of notation we consider $c_{s+1} = c_1$). Assume w.l.o.g. that $|c_{i+1} - c_i|$ is associated with a_i, for $1 \leq i \leq s$. We define $I^+ = \{i \in \{1, \ldots, s\} \mid c_{i+1} > c_i\}$ and $I^- = \{j \in \{1, \ldots, s\} \mid c_{j+1} < c_j\}$. This yields

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

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

3 NP-hardness of t-Partial Digest Superset

In the previous section, we proved NP-hardness for Min Partial Digest Superset by a reduction from Equal Sum Subsets. In the proof, we distinguished whether a minimal solution uses at most n points, or $n+1$ points. In this section, we will generalize this proof and allow to "shift" this boundary to some value t that is sufficiently large.

We will show that t-Partial Digest Superset is NP-hard for every $0 < \epsilon < 1/2$ if we set t to be at least $f(|D|) = |D|^{\frac{1}{2}+\epsilon}$. Observe that for a distance multiset D, a minimal set of points covering D has cardinality at least $\frac{1}{2}+\sqrt{\frac{1}{4}+2|D|} \approx |D|^{\frac{1}{2}}$. Moreover, the Partial Digest problem is equivalent to t-Partial Digest Superset with $t = \frac{1}{2}+\sqrt{\frac{1}{4}+2|D|} = O\left(|D|^{1/2}\right)$.

We need to introduce some notation for large numbers first. The numbers are expressed in the number system of some base Z. We denote by $\langle a_1, \ldots, a_n\rangle_Z$ the number $\sum_{1 \leq i \leq n} a_i Z^{n-i}$; we say that a_i is the i-th digit of this number. We will choose base Z large enough such that adding up numbers will not lead to carry-digits from one digit to the next. Therefore, we can add numbers digit by digit. The same holds for scalar products. 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 will drop the base Z from our notation if this is clear from the context. Moreover, we will allow different bases for each digit. We define the concatenation of two numbers by $\langle a_1, \ldots, a_n\rangle \odot \langle b_1, \ldots, b_m\rangle := \langle a_1, \ldots, a_n, b_1, \ldots, b_m\rangle$, i.e., $\alpha \odot \beta = \alpha Z^m + \beta$, 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 i-th 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 $\mathbb{1}_n = Z^n - 1$.

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

Proof (sketch). We will prove the theorem for the case $t = f(|D|)$, as the case $t \geq f(|D|)$ is a simple extension. In particular, we will show that Equal Sum Subsets can be reduced to $\leq_p$ t-Partial Digest Superset. Let $\{a_1, \ldots, a_n\}$ be an instance of Equal Sum Subsets. Informally speaking, we will "blow up" the instance of t-Partial Digest Superset used in the proof of Theorem 1, by first adding a set B of r "essential" distances with the property that any solution must use $r+1$ points to cover this set, and these points cannot be used for the distances in $\{a_1, \ldots, a_n\}$. Then, a suitable set C' of $O(r^2)$ "inessential" distances is used so to blow up the size of D; these distances are covered "for free" by the points used for B. Thus, $n+r$ points are sufficient if and only if the instance of Equal Sum Subsets is a YES instance.

We define the distances as numbers with base $Z = r^2 + \sum_{i=1}^{n} a_i$. Let $a_i' = \langle a_i\rangle \odot \mathbb{0}_r$ and $A' = \{a_i' \mid 1 \leq i \leq n\}$. For $1 \leq j \leq r$, let $b_j = \langle 0\rangle \odot \Delta_r(j)$, and $B = \{b_j \mid 1 \leq j \leq r\}$. For $1 \leq u < v \leq r$, let $c_{u,v} = \sum_{\ell=u}^{v} b_\ell$, and $C = \{c_{u,v} \mid 1 \leq u < v \leq r\}$.

The instance of t-Partial Digest Superset is defined as $D = A' \cup B \cup C'$, where $C' \subseteq C$. Clearly, $|D| = n + r + |C'|$ and $t = n + r$. We want $|C'|$ to satisfy $t = n + r = f(|D|) = (n + r + |C'|)^{1/2+\epsilon}$. To this aim it would suffice to take any C' with $|C'| = (n+r)^{\frac{2}{1+2\epsilon}} - (n+r)$. However, the latter number may not be an integer. In this case, the proof can be easily adjusted by considering $|C'| = \lfloor (n+r)^{\frac{2}{1+2\epsilon}} \rfloor - (n+r)$, a sufficiently large n, and r polynomial in n.

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 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 A'. Therefore, we need $r+1$ points to cover all distances from $B \cup C'$. The remaining set A' behaves as in the proof of Theorem 1: by reusing one of the $r+1$ points above, we we need at most n further points to cover A'; as in the proof of Theorem 1, less than n points are necessary if and only if there exists a solution for the Equal Sum Subsets instance. □

4 (In-) Approximability of Max Partial Digest Subset

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

We construct a reduction from Max Clique to Max Partial Digest Subset. Max Clique is the problem of finding a maximum complete subgraph from a given graph. It cannot be approximated by any polynomial-time algorithm with an approximation ratio of $n^{1-\epsilon}$ for any $\epsilon > 0$, where n is the number of vertices of the input graph, unless $\mathsf{P} = \mathsf{NP}$ [7]. Our reduction is gap-preserving, which means that the inapproximability of Max Clique is transfered to Max Partial Digest Subset.

Suppose we are given a graph $G = (V, E)$ 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 $d_{i,j} = \mathbb{0}_i \odot \mathbb{1}_{j-i} \odot \mathbb{0}_{n-j}$ with base $Z = n^2 + 1$ for each $(v_i, v_j) \in E, j > i$.

Let OPT be the size of the maximum clique in G (i.e., the number of vertices in the 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 $\epsilon > 0$. The following two lemmas show how the reduction works.

Lemma 3. $OPT \geq kn^{1-\epsilon} \implies OPT' \geq kn^{1-\epsilon}$

Proof. Assume we are given a clique in graph G of size $kn^{1-\epsilon}$. We construct a solution for the corresponding Maximum Partial Digest instance D by positioning a point at position $v'_i = \mathbb{1}_i \odot \mathbb{0}_{n-i}$ for each vertex v_i in the clique. This yields a feasible solution for D, since – for $j > i$ – each distance $v'_j - v'_i = \mathbb{0}_i \odot \mathbb{1}_{j-i} \odot \mathbb{0}_{n-j} = d_{i,j}$ 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. □

Lemma 4. $OPT < k \implies OPT' < k$

Proof. We prove the contraposition, i.e.,

$$OPT' \geq k \implies OPT \geq k.$$

Suppose we are given a solution of the Max Partial Digest Subset instance consisting of k points $p_1 < \ldots < p_k$ on the line, where 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 a single 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 either all of the form $\mathbb{0}_{i_{\min}} \odot \mathbb{1}_l \odot \mathbb{0}_{j_{\max}-l-i_{\min}} \odot \mathbb{0}_{n-j_{\max}}$ or all of the form $\mathbb{0}_{i_{\min}} \odot \mathbb{0}_l \odot \mathbb{1}_{j_{\max}-l-i_{\min}} \odot \mathbb{0}_{n-j_{\max}}$, where $i_{\min} \leq l \leq j_{\max}$. If both forms existed in a solution, 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 construct a vertex set V' that will turn out to be a clique by letting $v_{i_{\min}}$ and $v_{j_{\max}}$ be in this set V'. Additionally, for each $p_{k'}$ for $k' = 2, \ldots k-1$, where $p_{k'}$ is of the form $\mathbb{0}_{i_{\min}} \odot \mathbb{1}_{l'} \odot \mathbb{0}_{j_{\max}-l'-i_{\min}} \odot \mathbb{0}_{n-j_{\max}}$ or $\mathbb{0}_{i_{\min}} \odot \mathbb{0}_{l'} \odot \mathbb{1}_{j_{\max}-l'-i_{\min}} \odot \mathbb{0}_{n-j_{\max}}$, where $i_{\min} \leq l' \leq j_{\max}$, we let $v_{l'}$ be in the vertex set V'.

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 $v_{l'}$ into the set and $p_{k''}$ has led to the inclusion of vertex $v_{l''}$ into the clique. This difference is exactly $d_{l',l''}$ for both possible forms, and thus the edge $v_{l'}, v_{l''}$ 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 $kn^{1-\epsilon}$, or less than k, and we are to decide which is true, is NP-hard to decide [7]. Lemmas 3 and 4 transform 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 $kn^{1-\epsilon}$ or less than k points can be placed on a line. This promise problem of Max Partial Digest Subset is NP-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 $\mathsf{P} = \mathsf{NP}$, Maximum Partial Digest cannot be approximated with an approximation ratio of:

$$\frac{kn^{1-\epsilon}}{k} = n^{1-\epsilon} \geq |D|^{\frac{1}{2}-\epsilon},$$

where $|D|$ is the number of distances in instance D. We have shown the following:

Theorem 5. Max Partial Digest Subset *cannot be approximated by any polynomial-time algorithm with an approximation ratio of* $|D|^{\frac{1}{2}-\epsilon}$ *for any* $\epsilon > 0$, *where* $|D|$ *is the number of input distances, unless* $\mathsf{P} = \mathsf{NP}$.

A trivial approximation algorithm for a Max Partial Digest Subset instance $D = \{d_1, \ldots, d_{|D|}\}$ that simply places two points at distance d_1 from each other achieves a matching approximation ratio of $O(|D|^{\frac{1}{2}})$.

5 Conclusion and Open Problems

We have shown that the optimization problems Min Partial Digest Superset and Max Partial Digest Subset are NP-hard. Moreover, the maximization problem is not approximable within reasonable bounds, unless $\mathsf{P} = \mathsf{NP}$. This answers the problem left open in [11, Problem 12.116], and gives rise to new open questions:

1. 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 [16] seems to run in polynomial-time for real-life instances. How can this be explained? What relevant properties do real-life instances have that prevent them from becoming intractable?

2. What is the best approximation ratio for Min Partial Digest Superset?

3. Consider the following variation of Partial Digest: Given a *set* S of distances (instead of a multiset), find (a minimum/maximum number of) points on a line such that each distance between two of the points is in S. What is the computational complexity of this problem?

4. 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 are pairwise distinct?

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

Acknowledgments We would like to thank Dirk Bongartz and Walter Unger for pointing us to the Partial Digest problem, and Aris Pagourtzis and Peter Widmayer for their help in this work.

References

[1] F. Alizadeh, R. M. Karp, L. A. Newberg, and D. K. Weisser. Physical mapping of chromosomes: A combinatorial problem in molecular biology. In *Symposium on Discrete Algorithms*, pages 371–381, 1993.

[2] S. Arora and C. Lund. Hardness of approximations. In D. Hochbaum, editor, *Approximation Algorithms for NP-Hard Problems*, pages 399–446. PWS Publishing Company, 1996.

[3] 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.

[4] T. Dakić. *On the turnpike problem.* PhD thesis, Simon Fraser University, 2000.

[5] D. Fasulo. *Algorithms for DNA Restriction Mapping.* PhD thesis, University of Washington, 2000.

[6] J. Fütterer. Personal communication, 2002.

[7] J. Håstad. Clique is hard to approximate within $n^{1-\epsilon}$. In *Proc. of the Symposium on Foundations of Computer Science*, 1996.

[8] 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.

[9] L. Newberg and D. Naor. A lower bound on the number of solutions to the probed partial digest problem. *Advances in Applied Mathematics (ADVAM)*, 14:172–183, 1993.

[10] G. Pandurangan and H. Ramesh. The restriction mapping problem revisited. *Journal of Computer and System Sciences (JCSS)*, to appear 2002. Special issue on Computational Biology.

[11] P. Pevzner. *Computational Molecular Biology.* MIT Press, 2000.

[12] P. A. Pevzner and M. S. Waterman. Open combinatorial problems in computational molecular biology. In *Proc. of the Third Israel Symposium on Theory of Computing and Systems ISTCS*, pages 158–173. IEEE Computer Society Press, 1995.

[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 *Proceedings of the International IEEE Symposium on Intelligence in Neural and Biological Systems*, 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 *Sixth ACM Symposium on Computational Geometry*, 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] G. J. Woeginger and Z. L. Yu. On the equal-subset-sum problem. *Information Processing Letters*, 42:299–302, 1992.

[19] Z. Zhang. An Exponential Example for a Partial Digest Mapping Algorithm. *Journal of Computational Biology*, 1(3):235–239, 1994.