Noisy Data Make the Partial Digest Problem NP-hard*

Mark Cieliebak[1], Stephan Eidenbenz[2], and Paolo Penna[3]

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

Abstract. The problem to find 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 is known as PARTIAL DIGEST problem, which occurs for instance in DNA physical mapping and de novo sequencing of proteins. Although PARTIAL DIGEST was – as a combinatorial problem – already proposed in the 1930's, its computational complexity is still unknown.
In an effort to model real-life data, we introduce two optimization variations of PARTIAL DIGEST that model two different error types that occur in real-life data. First, 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. This 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 $|D|^{\frac{1}{2}-\varepsilon}$ for any $\varepsilon > 0$, where $|D|$ is the number of input distances. This inapproximability result is tight up to low-order terms as we give a trivial approximation algorithm that achieves a matching approximation ratio.

1 Introduction

The PARTIAL DIGEST problem is one of the most intriguing problems in computational biology: on the one hand, it is a basic problem with relevant applications in DNA mapping and in protein 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. More formally, the PARTIAL DIGEST problem can be defined as follows.

* A preliminary version of this paper has been published as Technical Report 381, ETH Zurich, Department of Computer Science, October 2002.
† LA-UR-03:1157; work done while at ETH Zurich.

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

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

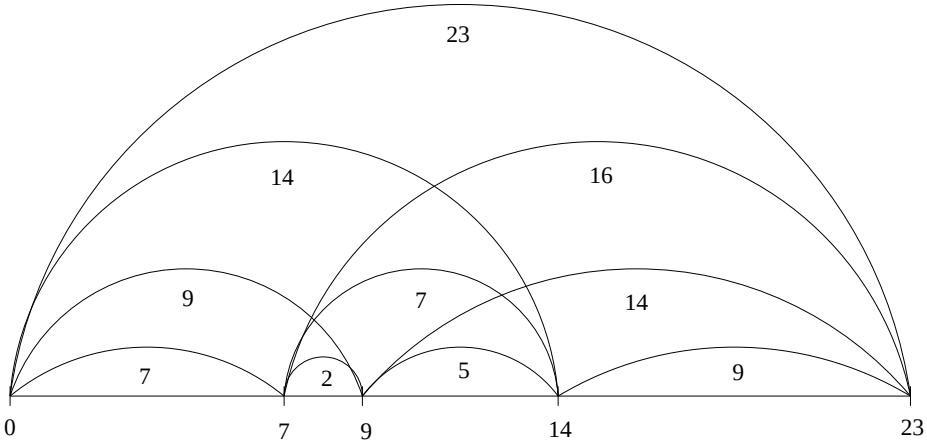

Fig. 1. Example for PARTIAL DIGEST

Recently, the PARTIAL DIGEST problem has received increasing attention due to its applications in computational biology, namely physical mapping of DNA and de novo sequencing of proteins (see below). However, in its pure combinatorial formulation the PARTIAL DIGEST problem has been studied for a long time: It appears already in the 1930's in the sphere of X-ray crystallography (acc. to [23]); the problem is very closely related to the theory of homometric sets[2] [20]; and finally, it 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 [8].

We refer to the problem as PARTIAL DIGEST due to it 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

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

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

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 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.

The PARTIAL DIGEST problem occurs as well in de novo sequencing of proteins using tandem mass spectrometry: Given a probe with many copies of a single protein, we first use an enzyme like trypsin to digest the proteins. This leaves us with a set of protein fragments, called peptides. We separate the peptides by their mass, using tandem mass spectrometry. Then we break up these peptides into even smaller fragments using collision induced dissociation (CID). The mass/charge ratio of these fragments are measured using mass spectrometry again, resulting in a *tandem mass spectrum* of the peptide, which can be used to determine the amino acid sequence of the peptide (*de novo sequencing*). 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 (e.g., peptide $AEKGCWTR$ can break up into fragments $AEKG$ and $CWRT$, or into fragments AE and $KGCWTR$), then only fragments occur that are prefixes and suffixes of the peptide sequence. In this case there exist efficient algorithms for de novo sequencing [6, 17]. However, in real life experiments a single peptide does not only break up once, but it can break up several times, yielding internal fragments as well [3, 4, 14]. In the example, peptide $AEKGCWTR$ might break up into fragments AEK, GC and WTR. For this reason, we do not only obtain prefixes and suffixes in the spectrum, but all possible substrings of the peptide sequencs. Hence, the problem to find the appropriate sequence of amino acid for such a spectrum is equivalent to the PARTIAL DIGEST problem.

For the sake of simplicity, we will refer in this paper only to the setup of partial digestion experiments for DNA molecules. It is obvious that similar types of noise occur in tandem mass spectrometry data as well.

In reality, the partial digest experiment cannot be conducted under ideal conditions as outlined above, and thus errors occur in the data. In fact, there are four types of errors [9, 11, 13, 24, 26]:

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 to it. Furthermore, fragments can be added through contamination with biological material, such as DNA from unrelated sources.

Missing fragments It can happen that a particular restriction site does not get cut in combination with all other restriction sites; then only one large fragment occurs in the data instead of the two (or even more) smaller fragments. Furthermore, fragments cannot be 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 brightness of its spot in the gel is a non–trivial problem.

In this paper, we define two optimization variations of PARTIAL DIGEST, where one variation models addition errors and the other models omission errors or missing fragments. 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. For the third type of errors, it is known that the PARTIAL DIGEST problem becomes NP-hard if length measurements are erroneous [7], while we are not aware of any results on multiplicity errors. Intuitively, the problem of modeling "real-life" instances – in which *all* error types can occur – is even harder than having only one error type.

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 2 (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 $D \subseteq \{|p_i - p_j| \mid 1 \leq i < j \leq m\}$.*

For example, if $D = \{2, 5, 7, 7, 9, 14, 23\}$, then the solution shown in Figure 1 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 shown in the figure would cover all distances from D', but there exist solutions with less points that cover D', e.g. points $P' = \{0, 2, 9, 18\}$ (yielding distance multiset $\{2, 7, 9, 9, 16, 18\}$).

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 3 (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 multiset 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-complete problem EQUAL SUM SUBSETS. In a sense, our result provides an answer to the open problem 12.116 in the book by Pevzner [18], which asks for an algorithm to reconstruct a set of points, given a subset of their pairwise distances. We 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; in this case, the problem remains NP-hard for *any* fixed $t = |D|^{\frac{1}{2}+\varepsilon}$ and any $\varepsilon > 0$. This result is tight in a sense, since any solution (even from the original PARTIAL DIGEST) must have at least cardinality $t = \Omega(|D|^{\frac{1}{2}})$. As for the MAX PARTIAL DIGEST SUBSET problem, we show that there is no polynomial–time algorithm for this problem that guarantees an approx-

imation ratio[3] of $|D|^{\frac{1}{2}-\varepsilon}$ for any $\varepsilon > 0$, unless $\mathsf{P} = \mathsf{NP}$, by proposing a gap-preserving reduction from MAXIMUM CLIQUE. The problem MAXIMUM CLIQUE is very hard to approximate, and our reduction transfers the inapproximability of MAXIMUM CLIQUE to MAX PARTIAL DIGEST SUBSET. We also point to a trivial approximation algorithm that achieves a matching asymptotic approximation ratio. Thus, our result is tight up to low-order terms. Our inapproximability result means not only that can we not expect a polynomial-time algorithm that finds the optimum solution, but we cannot even expect a polynomial-time algorithm for MAX PARTIAL DIGEST SUBSET that finds solutions that are a factor $|D|^{\frac{1}{2}-\varepsilon}$ off the optimum.

Our hardness results show that a polynomial-time algorithm for the original PARTIAL DIGEST (if any) cannot be obtained by looking at the 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 time[4] [15, 20]; there exists a backtracking algorithm (for exact or erroneous data) that has expected running time polynomial in the number of distances [23, 24], but exponential worst case running time [27]; 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 [21]; and finally, if the points are not on a line but in d-dimensional space, then the problem is NP-hard for some $d \geq 2$ [23]. However, for the original PARTIAL DIGEST problem, neither a polynomial–time algorithm nor a proof of NP-completeness is known [5, 8, 17–19, 22].

In the biological setting of partial digestion, many experimental variations have been studied: Double digestion, where two different enzymes are used [22]; probed partial digestion, where probes (markers) are hybridized to partially digested DNA [1, 16]; simplified partial digest, where clones are cleaved in either one or in all restriction sites [5]; labeled partial digestion, where both ends of the DNA molecule are labeled before digestion [17]; and multiple complete digestion, where many different enzymes

[3] 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.

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

are used [10]. For a good survey on the PARTIAL DIGEST problem, see [23]; and for more recent discussions on the problem, see [18] and [22].

The paper is organized as follows: In Section 2 we present the hardness results of MIN PARTIAL DIGEST SUPERSET. In Section 3 we provide the (in-) approximability results on MAX PARTIAL DIGEST SUBSET. Finally, we conclude and present some open problems in Section 4.

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., $\mathrm{sum}(X) := \sum_{x \in X} x$. Finally, we say that a set of points P *covers* distance multiset D if $D \subseteq \Delta(P)$.

We first show that the minimum cardinality of a point set that covers all distances in a given multiset D cannot be to large: 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 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 $k + 1$.

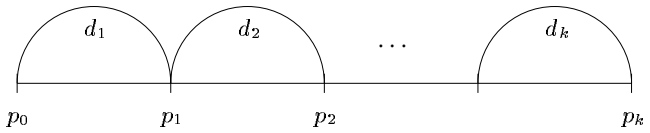

Fig. 2. Trivial solution for a distance multiset D.

Observe that PARTIAL DIGEST can be easily reduced to MIN PARTIAL DIGEST SUPERSET: Given an instance D of PARTIAL DIGEST of size $|D| = k$, there is a solution for D if and only if the minimal solution for the

MIN PARTIAL DIGEST SUPERSET instance D has size $m = \frac{1}{2} + \sqrt{\frac{1}{4} + 2k}$ (in this case, $k = \binom{m}{2}$).

Theorem 4. MIN PARTIAL DIGEST SUPERSET *is* NP*-hard.*

Proof. We reduce EQUAL SUM SUBSETS to MIN PARTIAL DIGEST SUPERSET, where EQUAL SUM SUBSETS is the NP-complete problem [25] 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 $\text{sum}(X) = \text{sum}(Y)$?

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 $1 \leq r < s \leq n$. 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 (cf. Figure 3); the remaining distances from $D - (X \cup Y)$ are at the end of the two chains.

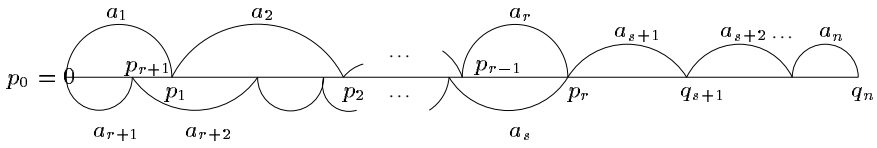

Fig. 3. Solution if there are two sets of equal sum.

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, by construction (cf. Figure 3), 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 $m < n + 1$, this graph contains a cycle. We show that such a cycle induces a solution of the EQUAL SUM SUBSETS instance.

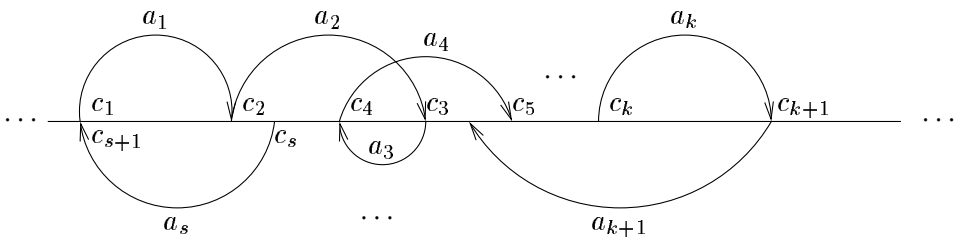

Fig. 4. A solution containing a cycle yields two subsets of equal sum: the overall lenght of right jumps equals to the overall length of left jumps.

Let $C = c_1, \ldots, c_s$ be a cycle in G (see Fig. 4). 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 $I^- = \{j \in \{1, \ldots, s\} \mid c_{j+1} < c_j\}$, 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

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

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. □

In the previous proof, we distinguished whether a minimal solution uses at most n points, or $n + 1$ points. It is even possible to "decrease" this boundary to some value t that is still sufficiently large. In fact, we

can show that t-PARTIAL DIGEST SUPERSET is NP-hard for every $0 < \varepsilon < \frac{1}{2}$ if we set t to be at least $f(|D|) = |D|^{\frac{1}{2}+\varepsilon}$. Observe that for a distance multiset D, a minimal set of points covering D has cardinality at least $\frac{1}{2} + \sqrt{\frac{1}{4} + 2|D|} \approx |D|^{\frac{1}{2}}$. Moreover, the PARTIAL DIGEST problem is equivalent to t-PARTIAL DIGEST SUPERSET with $t = \frac{1}{2} + \sqrt{\frac{1}{4} + 2|D|} = O\left(|D|^{\frac{1}{2}}\right)$. The proof will be given in the full version of this paper. It is a reduction from EQUAL SUM SUBSETS where we "blow up" the instance of MIN PARTIAL DIGEST SUPERSET used in the proof above by adding an appropriate number of additional distances that do not interfere.

3 (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 need to introduce some notation for large numbers first. The numbers are expressed in the number system of some base Z. We denote by $\langle a_1, \ldots, a_n \rangle$ the number $\sum_{1 \le i \le n} a_i Z^{n-i}$; 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 in our proof 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 allow different bases for each digit. 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 i-th position where the digit is 1. Moreover, $\mathbf{1}_n := \langle 1, \ldots, 1 \rangle$ has n digits, all 1's, and $\mathbf{0}_n := \langle 0, \ldots 0 \rangle$ has n zeros.

We construct a gap-preserving reduction (as introduced in [2]) 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 $n^{1-\varepsilon}$ for any $\varepsilon > 0$, where n is the number of vertices of the input graph, unless P = NP [12]. 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 $d_{i,j} = \mathbf{0}_i \circ \mathbf{1}_{j-i} \circ \mathbf{0}_{n-j}$ 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 $\varepsilon > 0$. The following two lemmas show how the reduction works.

Lemma 5. $OPT \geq kn^{1-\varepsilon} \implies OPT' \geq kn^{1-\varepsilon}$

Proof. Assume we are given a clique in graph G of size $kn^{1-\varepsilon}$. We construct a solution for the corresponding MAXIMUM PARTIAL DIGEST instance D by positioning a point at position $v'_i = \mathbf{1}_i \circ \mathbf{0}_{n-i}$ for each vertex v_i in the clique. This yields a feasible solution for D, since – for $j > i$ – each distance $v'_j - v'_i = \mathbf{0}_i \circ \mathbf{1}_{j-i} \circ \mathbf{0}_{n-j} = d_{i,j}$ 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 6. $OPT < k \implies OPT' < k$

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 $p_1 = \mathbf{0}_n$. By definition distance $p_k - p_1$ must be contained in the distance set D and thus two indices $i_{\min}$ and $j_{\max}$ must exist with $d_{i_{\min},j_{\max}} = p_k - p_1$. 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 $\mathbf{0}_{i_{\min}} \circ \mathbf{1}_l \circ \mathbf{0}_{j_{\max}-l-i_{\min}} \circ \mathbf{0}_{n-j_{\max}}$ or all of the form $\mathbf{0}_{i_{\min}} \circ \mathbf{0}_l \circ \mathbf{1}_{j_{\max}-l-i_{\min}} \circ \mathbf{0}_{n-j_{\max}}$, 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 $\mathbf{0}_{i_{\min}} \circ \mathbf{1}_{l'} \circ \mathbf{0}_{j_{\max}-l'-i_{\min}} \circ \mathbf{0}_{n-j_{\max}}$ or $\mathbf{0}_{i_{\min}} \circ \mathbf{0}_{l'} \circ \mathbf{1}_{j_{\max}-l'-i_{\min}} \circ \mathbf{0}_{n-j_{\max}}$, where $i_{\min} \leq l' \leq j_{\max}$, we let $v_{l'+i_{\min}}$ 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 $kn^{1-\varepsilon}$, or less than k, and we are to decide which is true, is NP-hard to decide [12]. Lemmas 5 and 6 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 $kn^{1-\varepsilon}$ 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.[5] Thus, unless P = NP, MAX PARTIAL DIGEST SUBSET cannot be approximated with an approximation ratio of:

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

where $|D|$ is the number of distances in instance D, since we could decide the corresponding promise problem in polynomial time otherwise. We have shown the following:

Theorem 7. MAX PARTIAL DIGEST SUBSET *cannot be approximated by any polynomial-time algorithm with an approximation ratio of* $|D|^{\frac{1}{2}-\varepsilon}$ *for any* $\varepsilon > 0$, *where* $|D|$ *is the number of input distances, unless* P = 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 $O(|D|^{\frac{1}{2}})$.

4 Conclusion and Open Problems

We have shown that the optimization problems MIN PARTIAL DIGEST SUPERSET and MAX PARTIAL DIGEST SUBSET are NP-hard. Moreover,

[5] The concept of gap-preserving reductions is an alternative way to view a reduction. It provides a formal framework for preserving inapproximability ratios between two optimization problems. For details, see [2].

the maximization problem is not approximable within reasonable bounds, unless P = NP. This answers the problem 12.116 left open in [18], 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 [23] 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. Using gel electrophoresis or mass spectrometry, 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 "to much" from the given multiplicities. 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 Sacha Baginsky, 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. V. Bafna and N. Edwards. On de novo interpretation of tandem mass spectra for peptide identification. In 7^{th} *Annual International Conference on Computational Biology (RECOMB 03)*, pages 9–18, 2003.

4. S. Baginsky. Personal communication, 2003.
5. 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.
6. T. Chen, M. Kao, M. Tepel, J. Rush, and G. M. Church. A dynamic programming approacht to de novo peptide sequencing via tandem mass spectrometry. In 11^{th} *SIAM-ACM Symposium on Discrete Algorithms (SODA)*, pages 389–398, 2000.
7. M. Cieliebak and S. Eidenbenz. Measurement errors make the partial digest problem NP-hard, manuscript, to be published, 2003.
8. T. Dakić. *On the turnpike problem.* PhD thesis, Simon Fraser University, 2000.
9. T. I. Dix and D. H. Kieronska. Errors between sites in restriction site mapping. *Computer Applications in the Biosciences (CABIOS)*, 4(1):117–123, 1988.
10. D. Fasulo. *Algorithms for DNA Restriction Mapping.* PhD thesis, University of Washington, 2000.
11. J. Fütterer. Personal communication, 2002.
12. J. Håstad. Clique is hard to approximate within $n^{1-\epsilon}$. In *Proc. of the Symposium on Foundations of Computer Science*, 1996.
13. J. Inglehart and P. C. Nelson. On the limitations of automated restriction mapping. *Computer Applications in the Biosciences (CABIOS)*, 10(3):249–261, 1994.
14. P. James. *Proteome Research: Mass Spectrometry.* Springer, 2001.
15. 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.
16. 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.
17. 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.
18. P. Pevzner. *Computational Molecular Biology.* MIT Press, 2000.
19. 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.
20. J. Rosenblatt and P. Seymour. The structure of homometric sets. *SIAM Journal of Algorithms and Discrete Mathematics*, 3(3):343–350, 1982.
21. D. B. Searls. Formal grammars for intermolecular structure. In *Proceedings of the International IEEE Symposium on Intelligence in Neural and Biological Systems*, 1995.
22. J. Setubal and J. Meidanis. *Introduction to Computational Molecular Biology.* PWS Boston, 1997.
23. S. S. Skiena, W. Smith, and P. Lemke. Reconstructing sets from interpoint distances. In *Sixth ACM Symposium on Computational Geometry*, pages 332–339, 1990.
24. S. S. Skiena and G. Sundaram. A partial digest approach to restriction site mapping. *Bulletin of Mathematical Biology*, 56:275–294, 1994.
25. G. J. Woeginger and Z. L. Yu. On the equal-subset-sum problem. *Information Processing Letters*, 42:299–302, 1992.

26. L. W. Wright, J. B. Lichter, J. Reinitz, M. A. Shifman, K. K. Kidd, and P. L. Miller. Computer-assisted restriction mapping: an integrated approach to handling experimental uncertainty. *Computer Applications in the Biosciences (CABIOS)*, 10(4):435–442, 1994.
27. Z. Zhang. An Exponential Example for a Partial Digest Mapping Algorithm. *Journal of Computational Biology*, 1(3):235–239, 1994.