On the Fixed Parameter Tractability and Approximability of the Minimum Error Correction Problem

Paola Bonizzoni, Riccardo Dondi, Gunnar W Klau, Yuri Pirola, Nadia Pisanti, Simone Zaccaria

On the Fixed Parameter Tractability and Approximability of the Minimum Error Correction problem

Paola Bonizzoni[1], Riccardo Dondi[2], Gunnar W. Klau[3,5], Yuri Pirola[1], Nadia Pisanti[4,5], and Simone Zaccaria[1]

[1] DISCo, Univ. degli Studi di Milano-Bicocca, Milan, Italy, {simone.zaccaria,bonizzoni,pirola}@disco.unimib.it
[2] Dip. di Scienze Umane e Sociali, Univ. degli Studi di Bergamo, Bergamo, Italy, riccardo.dondi@unibg.it
[3] Life Sciences, Centrum Wiskunde & Informatica (CWI), Amsterdam, The Netherlands, gunnar.klau@cwi.nl
[4] Dipartimento di Informatica, Univ. degli Studi di Pisa, Italy, pisanti@di.unipi.it
[5] Erable Team, INRIA

Abstract. Haplotype assembly is the computational problem of reconstructing the two parental copies, called *haplotypes*, of each chromosome starting from sequencing reads, called *fragments*, possibly affected by sequencing errors. *Minimum Error Correction* (MEC) is a prominent computational problem for haplotype assembly and, given a set of fragments, aims at reconstructing the two haplotypes by applying the minimum number of base corrections. By using novel combinatorial properties of MEC instances, we are able to provide new results on the fixed-parameter tractability and approximability of MEC. In particular, we show that MEC is in FPT when parameterized by the number of corrections, and, on "gapless" instances, it is in FPT also when parameterized by the length of the fragments, whereas the result known in literature forces the reconstruction of complementary haplotypes. Then, we show that MEC cannot be approximated within any constant factor while it is approximable within factor $O(\log nm)$ where nm is the size of the input. Finally, we provide a practical 2-approximation algorithm for the Binary MEC, a variant of MEC that has been applied in the framework of clustering binary data.

1 Introduction

The genome of diploid organisms, as humans, is composed of two parental copies, called *haplotypes*, for each chromosome. The most frequent form of genetic variations between the two haplotypes of the same chromosome are the *Single Nucleotide Polymorphisms (SNPs)*. Haplotype analysis is of fundamental importance for a variety of applications including medical diagnostic and drug design [3, 4, 18].

The task of the *haplotyping problem* is the reconstruction of each pair of haplotypes. However, large scale direct experimental reconstruction from the collected samples is not yet cost-effective. One of the computational approaches that have been proposed, *haplotype assembly*, considers the high-throughput sequencing reads (also called *fragments*) that have to be bipartitioned in order to reconstruct the two haplotypes. Since for most of the SNP positions only two nucleotides are seen, the haplotypes can be represented as binary vectors. The fragments obtained from sequencing may not cover some positions of the haplotypes. These uncovered positions are called *holes*, whereas a sequence of holes within a fragment is called *gap*. However, the presence of sequencing and (possible) mapping errors makes the haplotype assembly problem a challenging task. In literature, different combinatorial formulations of the problem have been proposed [1, 7, 17, 18]. Among them, *Minimum Error Correction* (MEC) [18] has been proved particularly successful in the reconstruction of accurate haplotypes [5, 13, 20]. However, MEC is a computationally hard problem. Indeed, MEC is APX-hard even if the fragments have at least one gap [6] and remains NP-hard even if the fragments do not contain gaps (*Gapless MEC*) [6]. Instead, the computational complexity of MEC on instances without holes – called *Binary MEC* – is still unknown. Many successful approaches for coping with the computational intractability of MEC are based on the parameterized complexity framework. In particular, MEC is in FPT when parameterized by the "coverage" [20], that is the maximum number of fragments with non-hole values on a SNP position. Moreover, MEC is in FPT also when parameterized by the length of the fragments [13], but only under the *all-heterozygous assumption*, that forces to reconstruct complementary haplotypes.

Despite the significant amount of work present in the literature, some important questions related to the fixed-parameter tractability and approximability of MEC are still open. Two significant open problems are whether there exists a constant approximation algorithm for MEC and whether MEC is in FPT when parameterized by parameters of classical or practical interest, such as the total number of corrections or the length of the fragments. Indeed, removing the dependency on the all-heterozygous assumption from [13] does not appear straightforward and, hence, fixed-parameter tractability of MEC when parameterized by the fragment length is still an open problem.

The binary restriction of MEC where the fragments do not contain holes is particularly interesting from a mathematical point of view, and is the variant of the well-known *Hamming k-Median Clustering Problem* [6, 16], when $k = 2$. This clustering problem asks for k representative "consensus" (also called "median") strings with the goal of minimizing the distance between each input string and its closest consensus string. Hamming 2-Median Clustering is well studied from the approximation viewpoint, and a Polynomial Time Approximation Scheme (PTAS) has been proposed, both in a randomized [19] and deterministic form [14].

In this work, we present advances in the characterization of the fixed-parameter tractability and the approximability of MEC problem in the general, gapless, and binary cases. We first show that MEC is not in APX, *i.e.*, it is not approximable within constant factor. However, we also show that a reduction previously known [8] can be adapted to prove that MEC is approximable within factor $O(\log nm)$ (where n is the number of fragments and m is the number of SNPs) and that MEC is in FPT when parameterized by the total number of corrections.

Furthermore, by inspecting novel combinatorial properties of gapless instances, we show that Gapless MEC is in FPT when parameterized by the length of the fragments and that Binary MEC can be approximated within factor 2. Although

Binary MEC is known to admit a PTAS, the 2-approximation algorithm we give is more practical and intuitive than the previous approximation results.

2 Preliminary Definitions

In this section, we introduce some basic notions and the formal definition of the MEC problem. In the rest of the work, we indicate, as usual, the value of a vector s at position t as $s[t]$.

A *fragment matrix* is a matrix $\mathcal{M}$ composed of n rows and m columns such that each entry contains a value in $\{0, 1, -\}$. Each row of $\mathcal{M}$ represents a *fragment* and, formally, is a vector belonging to $\{0, 1, -\}^m$. Symmetrically, each column of $\mathcal{M}$ corresponds to a SNP position and is a vector belonging to $\{0, 1, -\}^n$. We denote by f_i the i-th row of $\mathcal{M}$ and by p_j the j-th column of $\mathcal{M}$. As a consequence, the entry of $\mathcal{M}$ at the i-th row and j-th column is denoted by $f_i[j]$ or $p_j[i]$. The *length* ℓ_i of a fragment f_i is defined as the number of elements in f_i between the rightmost and the leftmost non-hole elements (included) and we denote by ℓ the maximum length over all the fragments in $\mathcal{M}$. Moreover, we say that a column p_j *covers* a row f_i if $p_j[i] \in \{0, 1\}$. A column p_j is *heterozygous* if it contains both 0's and 1's, otherwise is *homozygous*. A *hole* is an entry $f_i[j]$ of $\mathcal{M}$ equal to the symbol $-$. A *gap* in a fragment f_i is a maximal subvector of holes in f_i surrounded by non-hole entries (that is, there exist two positions j_1 and j_2 with $j_1 + 1 < j_2$ such that $f_i[j_1], f_i[j_2] \neq -$ and $f_i[t] = -$ for all t with $j_1 < t < j_2$). A fragment matrix is *gapless* if no fragment contains a gap.

Two rows f_{i_1} and f_{i_2} are in *conflict* when there exists a position j, with $1 \leq j \leq m$, such that $f_{i_1}[j] \neq f_{i_2}[j]$, and $f_{i_1}[j], f_{i_2}[j] \neq -$. Otherwise, we say that f_{i_1} and f_{i_2} are in *agreement*. A collection $\mathcal{F}$ of fragments is in *agreement* if any pair of fragments f_1, f_2 in $\mathcal{F}$ are in agreement. A fragment matrix $\mathcal{M}$ is *conflict free* if there exists a bipartition $(\mathcal{F}_1, \mathcal{F}_2)$ of its fragments such that both $\mathcal{F}_1$ and $\mathcal{F}_2$ are in agreement.

When a fragment matrix $\mathcal{M}$ is conflict free, all the fragments in each part of the bipartition can be merged in order to reconstruct a haplotype, intended as a fragment without holes. Unfortunately, a fragment matrix $\mathcal{M}$ is not always conflict free. The Minimum Error Correction problem deals precisely with this issue by asking for a minimum set of *corrections* that make a fragment matrix conflict free, where a correction of a given fragment f_i at position j, with $f_i[j] \neq -$, is the flip of the value $f_i[j]$, replacing a 0 with a 1 or a 1 with a 0.

Problem 1 (Minimum Error Correction (MEC) problem).
Input: a matrix $\mathcal{M}$ of n rows and m columns.
Output: a conflict free matrix $\mathcal{M}'$ obtained from $\mathcal{M}$ with the minimum number of corrections.

Gapless MEC is the restriction of MEC where the input fragment matrix $\mathcal{M}$ is gapless, while *Binary MEC* is the restriction of (Gapless) MEC where the input fragment matrix $\mathcal{M}$ does not contain holes (that is, when $\mathcal{M}$ is a binary matrix).

Given a conflict free fragment matrix $\mathcal{M}$, any heterozygous column p_j encodes a bipartition of the fragments covered by p_j indicating which one belongs to one haplotype and which one belongs to other. Instead, any homozygous column p_j gives no information on how the covered fragments have to be partitioned, and it is "in accordance" with any other bipartition or heterozygous column. More formally, we say that two columns p_{j_1}, p_{j_2} of a fragment matrix are in *accordance* if (1) at least one of p_{j_1}, p_{j_2} is homozygous or (2) p_{j_1}, p_{j_2} are both heterozygous and are identical or complementary on the fragments covered by both.

As stated in the following property, that can be easily proved by induction, pairwise column accordance on gapless matrices is a necessary and sufficient condition for being conflict free.

Property 2. Let $\mathcal{M}$ be a *gapless* fragment matrix. Then, $\mathcal{M}$ is conflict free if and only if each pair of columns is in accordance.

Such a property is particularly important when designing exact algorithms for Gapless MEC, as it allows to test only for pairwise column accordance in order to ensure that the matrix is conflict free. In fact, the fixed-parameter algorithm for Gapless MEC that we present in Sect. 4 is based on this property. Furthermore, notice that if we relax the requirement that $\mathcal{M}$ is gapless, then the property does not hold. Consider, for example, the fragment matrix $\mathcal{M}$ composed of three fragments $f_1 = 01-$, $f_2 = -01$, and $f_3 = 1-0$. The three columns are pairwise in accordance, but the matrix is not conflict free (and, in fact, f_3 contains a gap).

Given two columns p_{j_1}, p_{j_2} of a fragment matrix $\mathcal{M}$, we define their (generalized) Hamming distance $d_H(p_{j_1}, p_{j_2})$ as $|\{i \mid \{p_{j_1}[i], p_{j_2}[i]\} = \{0,1\}\}|$ while their *correction distance* $d(p_{j_1}, p_{j_2})$ as the minimum between $d_H(p_{j_1}, p_{j_2})$ and $d_H(\overline{p_{j_1}}, p_{j_2})$ (where $\overline{p}$ is the complement of p on non-hole entries). Notice that the correction distance is non-negative and symmetric, but does not satisfy the triangle inequality, hence, despite the name, is not a metric. We also define the *homozygous distance* $H(p_j)$ as the minimum between the number of 0's and 1's contained in p_j. Intuitively, the correction distance is the cost of making a column equal or complementary to another column, while the homozygous distance is the cost of making a column homozygous.

A solution of MEC over a fragment matrix $\mathcal{M}$ is a bipartition of its fragments, that can be encoded as a binary vector O. It is easy to see that the cost of that solution is: $cost_{\mathcal{M}}(O) = \sum_{j=1}^{m} \min(d(O, p_j), H(p_j))$.

3 Inapproximability of MEC

In this section, we show that MEC is not in APX, that is MEC cannot be approximated within constant factor. We achieve this result by introducing an L-reduction from the Edge Bipartization problem to MEC.

The Edge Bipartization problem is defined as follows.

Problem 3 (Edge Bipartization (EB) problem [9]).
Input: an undirected graph $G = (V, E)$.
Output: $E' \subseteq E$ of minimum size such that $G' = (V, E \setminus E')$ is bipartite.

Now, we present the details of the reduction. Given an undirected graph $G = (V, E)$, we build the associated fragment matrix $\mathcal{M}(G)$ (with $|V|$ rows and $|E|$ columns) by setting, at each column p_j associated with edge $e_j = (u, v) \in E$, $f_u[j] = 0$, $f_v[j] = 1$, and $f_z[j] = -$ for $z \neq u, v$. Notice that, by construction, there exists a conflict in $\mathcal{M}(G)$ between fragments f_u and f_v if and only if $(u, v) \in E$.

Lemma 4. *Let $G = (V, E)$ be an undirected graph and $\mathcal{M}(G)$ be the associated fragment matrix. Given a solution E' of EB over G, we can compute a solution of MEC over $\mathcal{M}(G)$ with $|E'|$ corrections. Symmetrically, given a solution of MEC over $\mathcal{M}(G)$ with h corrections, we can compute a solution E' of EB over G of size at most h.*

Proof. ($\Rightarrow$) Let E' be a set of edges such that $(V_1 \uplus V_2, E \setminus E')$ is bipartite, where V_1 and V_2 are the parts of the bipartition. Build a matrix $\mathcal{M}'(G)$ from $\mathcal{M}(G)$ by flipping, for each $e_j = (u, v) \in E'$, the entry $f_u[j]$. Clearly, $\mathcal{M}'(G)$ is obtained from $\mathcal{M}(G)$ with $|E'|$ corrections and it does not contain conflicts induced by edges in E'. Let $(\mathcal{F}_1, \mathcal{F}_2)$ be the bipartition of fragments of $\mathcal{M}'(G)$ such that $\mathcal{F}_i := \{f_u \mid v_u \in V_i\}$ (for $i \in \{1, 2\}$). Each $\mathcal{F}_i$ is in agreement because it does not contain a pair of fragments associated with the endpoints of an edge of $E \setminus E'$. Hence, $\mathcal{M}'(G)$ is conflict free.

($\Leftarrow$) Let $\mathcal{M}'(G)$ be a conflict free matrix obtained from $\mathcal{M}(G)$ with h corrections and let C' be the subset of columns of $\mathcal{M}'(G)$ that contain exactly one correction. Consider the set $E' := \{e_j \in E \mid p_j \in C'\}$. Clearly, $|E'| \leq h$. Since $\mathcal{M}'(G)$ is conflict free, there exists a bipartition $(\mathcal{F}_1, \mathcal{F}_2)$ of the fragments such that both $\mathcal{F}_1, \mathcal{F}_2$ are in agreement. Build sets V_1, V_2 such that $V_i := \{v_u \mid f_u \in \mathcal{F}_i\}$ (with $i \in \{1, 2\}$). We claim that $(V_1 \uplus V_2, E \setminus E')$ is bipartite. Suppose to the contrary that there exists an edge $e_j = (u, v) \in E \setminus E'$ such that $u, v \in V_i$, $i \in \{1, 2\}$. Then, $f_u[j] = f_v[j]$ in $\mathcal{M}'(G)$, which implies that exactly one of $f_u[j]$ and $f_v[j]$ has been corrected (since $f_u[j] \neq f_v[j]$ in $\mathcal{M}(G)$). As a consequence, we have that $e_j \in E'$, contradicting the assumption.

Khot [15] proved that, under the Unique Games Conjecture, EB is not in APX. Since Lemma 4 proves that MEC is L-reducible to EB, we have the following result.

Theorem 5. *Under the Unique Games Conjecture [15], MEC is not in* APX.

The inapproximability result given in Theorem 5 nicely complements an approximation (and fixed-parameter tractable) result that can be easily inferred by a reduction presented in [8]. In [8], MEC is reduced to the Maximum Bipartite Induced Subgraph problem (MBIS). Given a vertex-weighted graph G, MBIS asks for a maximum weight subset of vertices of G that induces a bipartite graph. The reduction defines a graph, called *fragment graph*, whose nodes is the union of two sets: a set of nodes, called *fragment nodes*, one for each fragment, and a set of nodes, called *entry nodes*, one for each entry of the matrix. In order to avoid the removal of fragments nodes, they are assigned a sufficiently large weight.

The reduction can be easily reworked in order to prove approximation and fixed-parameter tractability results for MEC. More precisely, MEC is now reduced to the *Graph Bipartization* (GB) problem, a problem related to MBIS. Given an unweighted graph G, GB asks for the minimum number of vertex removals so that the resulting graph is bipartite. The reduction given in [8] can be modified by defining a new version of the fragment graph (see Fig. 1), where each (weighted) fragment node is substituted with a sufficiently large set of fragment nodes. From the construction of the fragment graph, it follows that a fragment matrix $\mathcal{M}$ is conflict free if and only if the corresponding fragment graph is bipartite and that a solution of MEC with k corrections corresponds to a solution of GB that removes k vertices.

	p_1	p_2	p_3	p_4
f_1	1	0	-	1
f_2	-	1	0	0
f_3	0	-	1	1

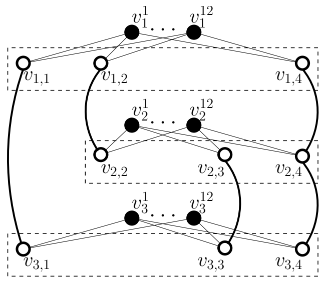

Fig. 1. A 3×4 fragment matrix (left) and the associated *fragment graph* (right). Fragment-nodes are in black, while entry-nodes are in white.

Since GB can be approximated within factor $O(\log |V|)$ [10] and is in FPT when parameterized by the number of removed vertices [11, 21], we have that:

Theorem 6. *(1) MEC can be approximated in polynomial time within factor $O(\log nm)$ where n is the number of fragments and m is the number of SNP positions.*
(2) MEC is in FPT when parameterized by the total number of corrections.

4 Gapless MEC is in FPT when parameterized by the fragment length

In this section, we introduce a fixed-parameter tractable algorithm for Gapless MEC when parameterized by the maximum length ℓ of the fragments. The algorithm is based on a dynamic programming approach and aims at finding a specific tripartition for the columns of a gapless fragment matrix $\mathcal{M}$. Due to space constraints, proofs of the lemmas of this section are moved to the appendix. In this section, we assume w.l.o.g. that $\mathcal{M}$ is a gapless fragment matrix and the fragments of $\mathcal{M}$ are sorted by starting position.

Firstly, we show a result that directly derives from Property 2. The following proposition stresses the relationship between a bipartition of the fragments and a tripartition of the columns in a gapless fragment matrix $\mathcal{M}$ that is conflict free.

Proposition 7. *Given a gapless fragment matrix $\mathcal{M}$, the following assertions are equivalent:*

1. *$\mathcal{M}$ is conflict free.*
2. *There exists a bipartition $(\mathcal{F}_1, \mathcal{F}_2)$ of the fragments, where both $\mathcal{F}_1$ and $\mathcal{F}_2$ are in agreement.*
3. *There exists a tripartition $T = (L, H, R)$ of the columns such that each column in H is homozygous, each column in $L \cup R$ is heterozygous, $d_H(p_{j_1}, p_{j_2}) = 0$ for all the columns $p_{j_1}, p_{j_2} \in L$ ($p_{j_1}, p_{j_2} \in R$, resp.) and $d_H(\overline{p_{j_1}}, p_{j_2}) = 0$ for each column $p_{j_1} \in L$ and each column $p_{j_2} \in R$.*

Based on Proposition 7, we introduce an algorithm for Gapless MEC that builds a tripartition of the columns of $\mathcal{M}$ in order to find a conflict free matrix $\mathcal{M}'$ obtained from $\mathcal{M}$ with the minimum number of corrections. Notice that in the rest of this

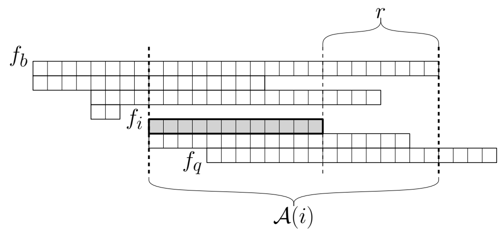

Fig. 2. The set $\mathcal{A}(i)$ of active columns for a fragment f_i.

section we implicitly refer only to tripartitions built as reported in the third assertion of Proposition 7.

The algorithm iteratively proceeds row-wise and, at each step, computes a tripartition for the columns considered so far. In particular, the key observation that allows to bound the exponential complexity of the algorithm to the parameter ℓ is that we can build any tripartition for all the columns in $\mathcal{M}$ by adding only a subset of columns, called *active columns*, for each row. This subset contains the columns covering the current fragment and the columns covering both previous and successive fragments. Indeed, we need to remember the tripartition established by previous fragments for columns that are covered by successive fragments. More formally, we define the set *active columns* for a fragment f_i as:

$$\mathcal{A}(i) = \{p_j \mid (p_j[i] \neq -) \vee (\exists x, y \text{ with } x < i < y \mid p_j[x], p_j[y] \neq -)\}$$

Fig. 2 represents the active columns $\mathcal{A}(i)$ of a fragment f_i. The cardinality of $\mathcal{A}(i)$ is bounded by ℓ. In fact, considering a row f_i, notice that $\ell_i \leq \ell$ and no column p_k, to the left of f_i, is in $\mathcal{A}(i)$. Assume that r is the number of columns p_j to the right of f_i, such that there are f_b, f_q with $b < i < q$ and $p_j[b], p_j[q] \neq -$. Since the r columns must be contained in $\mathcal{A}(b)$ for a fragment f_b with a starting position preceding the one of f_i, it holds that $\ell_i + r \leq \ell_b \leq \ell$. It clearly follows that $|\mathcal{A}(i)| = \ell_i + r \leq \ell$.

Considering two rows f_{i_1} and f_{i_2}, with $i_1 < i_2$, a tripartition for all the columns in $\mathcal{A}(i_1) \cup \mathcal{A}(i_2)$ can be computed whether there is a tripartition T_1 for $\mathcal{A}(i_1)$ and a tripartition T_2 for $\mathcal{A}(i_2)$, such that T_1 and T_2 are "in accordance", that is partitioning the shared columns in the same way. For this reason, we say that a tripartition $T_2 = (L_2, H_2, R_2)$ for $\mathcal{A}(i_2)$ *extends* another tripartition $T_1 = (L_1, H_1, R_1)$ for $\mathcal{A}(i_1)$ if and only if $L_1 \cap \mathcal{A}(i_2) \subseteq L_2$, $H_1 \cap \mathcal{A}(i_2) \subseteq H_2$, and $R_1 \cap \mathcal{A}(i_2) \subseteq R_2$.

At each step i, the algorithm computes a tripartition T for $\mathcal{A}(i)$ extending a tripartition T' for $\mathcal{A}(i-1)$. Since, $\mathcal{A}(i-1)$ also contains all the columns p_j with $p_j[i-1] = -$ such that there exists $y < i-1$ with $p_j[y] \neq -$ and $p_j[i] \neq -$, it follows that T even extends any tripartition computed at the previous steps extended by T'. As a consequence, we prove the following implication.

Lemma 8. *If there exists a conflict free matrix $\mathcal{M}''$ obtained from $\mathcal{M}$ on the first $i-1$ rows that induces a tripartition T' for the columns in $\mathcal{A}(i-1)$, and if T is a tripartition for the columns in $\mathcal{A}(i)$ extending T', then there exists a conflict free matrix $\mathcal{M}'$ obtained from $\mathcal{M}$ on the first i rows that induces the tripartition T for the columns in $\mathcal{A}(i)$.*

At each step i and for each tripartition $T = (L, H, R)$ for $\mathcal{A}(i)$, the algorithm chooses the tripartition T' extended by T for $\mathcal{A}(i-1)$ that induces the minimum cost (*recursive step*) and computes the minimum number of corrections to add on the current fragment f_i in order to tripartite all the columns in $\mathcal{A}(i)$ according to T (*local contribution*). In particular, the algorithm considers the minimum number of corrections on f_i such that $p_j[i] = 1$ or $p_j[i] = 0$ for all p_j in L and, on the contrary, $p_j[i] = 0$ or $p_j[i] = 1$ for all p_j in R. At the same time, the minimum number of corrections on the fragment f_i is computed for each column p_j in H such that p_j on the first i rows can be optimally transformed into a homozygous column. Therefore, we define $D[i, T]$ as the minimum number of corrections to obtain a conflict free matrix $\mathcal{M}'$ from $\mathcal{M}$ on the first i rows that induces a tripartition T for $\mathcal{A}(i)$. The algorithm proceeds row-wise computing the value $D[i, T]$ for each fragment f_i and for each tripartition T for $\mathcal{A}(i)$ by the following recursive equation:

$$D[i, T] = \Delta(i, T) + \min_{T' \text{ extended by } T} D[i-1, T'] \tag{1}$$

where T' is a tripartition for $\mathcal{A}(i-1)$. In the recursion, we consider only the tripartitions T' extended by T, since the shared columns have to be partitioned in the same way. In conclusion, the local contribution is defined as:

$$\Delta(i, T) = O(i, H) + \min \begin{cases} E^0(i, L) + E^1(i, R) \\ E^1(i, L) + E^0(i, R) \end{cases} \quad \textit{where } T = (L, H, R) \tag{2}$$

such that $E^x(i, F)$ is the cost of correcting equal to x the columns in F for fragment f_i, that is $E^x(i, F) = |\{j \mid j \in F \wedge p_j[i] \notin \{x, -\}\}|$, and $O(i, H)$ is the minimum number of corrections to apply on fragment f_i such that the columns in H, considered on the first i rows, can be turned into homozygous columns with minimum cost. Denote by $\#^x_{i,j}$ the number of values equal to x in $\{p_j[1], \ldots, p_j[i]\}$. The minimum between $\#^0_{i,j}$ and $\#^1_{i,j}$ states the minimum number of corrections necessary to turn a column p_j on the first i rows into a homozygous column. Since $O(i, H)$ refers only to the corrections on fragment f_i, we can compute $O(i, H)$ as:

$$O(i, H) = \sum_{j \in H} \begin{cases} 1 & p_j[i] = 0 \text{ and } \#^0_{i,j} \leq \#^1_{i,j} \\ 1 & p_j[i] = 1 \text{ and } \#^1_{i,j} \leq \#^0_{i,j} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Given a set of columns F, it is easy to see that $\sum_{i \in \{1,\ldots,n\}} O(i, F) = \sum_{p_j \in F} H(p_j)$.

The base case of the recurrence is $D[1, T] = \Delta(1, T)$ for each tripartition T for $\mathcal{A}(1)$. The algorithm returns the optimum corresponding to $\min_T D[n, T]$ where T is a tripartition for $\mathcal{A}(n)$. Furthermore, an optimal tripartition for all the columns can be computed by backtracking.

The algorithm computes all the values $D[i, T]$ for each tripartition T of the columns in $\mathcal{A}(i)$ and for each i in $\{1, \ldots, n\}$. It follows that there are $O(3^\ell \cdot n)$ entries and, therefore, the space complexity is equal to $O(3^\ell \cdot n)$. Given a tripartition T, we need $O(3^\ell)$ time to enumerate all the tripartitions T' extended by T because we have to tripartite all the columns in $|\mathcal{A}(i-1) \setminus \mathcal{A}(i)|$ with $\mathcal{A}(i-1) \leq \ell$ and, consequently, $|\mathcal{A}(i-1) \setminus \mathcal{A}(i)| \leq \ell$. Since $\Delta(i, T)$ can be computed in $O(\ell)$ time, each entry $D[i,T]$ can be computed in $O(3^\ell \cdot \ell)$. It follows that the total running time of the algorithm is $O(3^{2\ell} \cdot \ell \cdot n)$. Notice that storing partial information during the computation we can decrease the complexity to $O(3^\ell \cdot \ell \cdot n)$.

We now show the correctness of the algorithm.

Lemma 9. *Consider a gapless fragment matrix $\mathcal{M}$.*

1. *If $D[i,T] = h$, then there exists a conflict free matrix $\mathcal{M}'$ obtained from $\mathcal{M}$ on the first i rows with h corrections that induces a tripartition T for the columns in $\mathcal{A}(i)$.*
2. *If $\mathcal{M}'$ is a conflict free matrix obtained from $\mathcal{M}$ on the first i rows with h corrections that induces a tripartition T for the columns in $\mathcal{A}(i)$, then $D[i,T] \leq h$.*

From the analysis of the time complexity and the correctness of the algorithm, it directly follows that:

Theorem 10. *Gapless MEC is in FPT when parameterized by the length of the fragments.*

5 A 2-approximation algorithm for Binary MEC

In this section we present a 2-approximation algorithm for Binary MEC, that is the restriction of MEC where the fragment matrix does not contain holes. The approximation algorithm is based on the observation that heterozygous columns in binary matrices naturally encode bipartitions of the fragments and that, by Property 2, if the columns of a gapless fragment matrix are pairwise in accordance then the matrix is conflict free. In particular, Algorithm 1 builds a feasible solution $\texttt{SOL}[t]$ for each t in $\{1, \ldots, m\}$ assuming that p_t is the closest column to an (unknown) optimal bipartition O of the fragments. Each solution $\texttt{SOL}[t]$ corrects columns $p_{j'}$ with $H(p_{j'}) \leq d(p_t, p_{j'})$ into homozygous columns (equal to $\underline{1}$ or $\underline{0}$ depending on best choice), whereas it corrects the remaining columns $p_{j''}$ with $d(p_t, p_{j''}) < H(p_{j''})$ into heterozygous columns equal (or complementary, depending on the best choice) to p_t. It is easy to see that $\texttt{SOL}[t]$ for each t in $\{1, \ldots, m\}$ is a feasible solution (by Property 2) and that its cost is exactly $cost_{\mathcal{M}}(p_t)$.

Algorithm 1 A 2-approximation algorithm for Binary MEC

Require: A $n \times m$ binary matrix $\mathcal{M}$
  for $t = 1$ **to** m **do**   ▷ Assume that p_t is the column "closest" to O
    for $j = 1$ **to** m **do**
      if $H(p_j) \leq d(p_t, p_j)$ **then**
        Set p_j homozygous in $\texttt{SOL}[t]$
      else
        Set p_j equal/complementary to p_t in $\texttt{SOL}[t]$
  return $\arg\min_{\texttt{SOL}[t]} cost_{\mathcal{M}}(p_t)$

Algorithm 1 is a 2-approximation algorithm for Binary MEC.

Lemma 11. *Given a fragment matrix $\mathcal{M}$ without holes, if OPT is the optimum for Binary MEC on input $\mathcal{M}$, then Algorithm 1 returns a feasible solution with cost OPT' such that $OPT' \leq 2 \cdot OPT$.*

Proof. Assume that p_O is the column of $\mathcal{M}$ closest to an optimal bipartition O, that is $d(O,p_O) \leq d(O,p_j)$ for each j in $\{1,\ldots,m\}$ and assume that $d_H(O,p_O) \leq d_H(\overline{O},p_O)$ (if $d_H(\overline{O},p_O) < d_H(O,p_O)$ we can substitute O with $\overline{O}$ since they encode the same bipartition). Clearly, one such a column exists and $d_H(O,p_O) \leq d(O,p_j)$ for each j in $\{1,\ldots,m\}$. We show that, under this assumption, $d(p_O,p_j) \leq 2d(O,p_j)$. By the triangle inequality, $d_H(p_O,p_j) \leq d_H(p_O,O) + d_H(O,p_j)$. Hence, since $d_H(p_O,O) \leq d(O,p_j) \leq d_H(O,p_j)$, we have $d_H(p_O,p_j) \leq 2d_H(O,p_j)$. Similarly, we can prove that $d_H(p_O,\overline{p_j}) \leq 2d_H(O,\overline{p_j})$. As a consequence we have that $d(p_O,p_j) \leq 2d_H(O,p_j)$ and that $d(p_O,p_j) \leq 2d_H(O,\overline{p_j})$, which then imply $d(p_O,p_j) \leq 2d(O,p_j)$. Clearly, since $d(p_O,p_j) \leq 2d(O,p_j)$, we also have that $\min(d(p_O,p_j),H(p_j)) \leq 2\min(d(O,p_j),H(p_j))$.

Since Algorithm 1 considers all the columns as possible closest column to the unknown optimal bipartition O, we have that the cost of the returned solution is $OPT' \leq cost_{\mathcal{M}}(p_O) \leq 2\min(d(O,p_j),H(p_j)) = 2OPT$.

The running time of Algorithm 1 is $O(m^2 n)$ and, due to its simplicity, it is a more direct and practical approach than the PTAS algorithms known in literature [14,19].

6 Conclusions

Minimum Error Correction is a prominent combinatorial problem for haplotype assembly. Investigating the approximation complexity and the fixed-parameter tractability of MEC has proven useful to develop practical haplotype assembly tools [2,13,20]. Despite in this paper we addressed some issues that were left open, some other theoretical questions still need an answer.

In this work, we showed that, under the Unique Games Conjecture, MEC is not approximable within any constant factor. However, the approximation complexity of Gapless MEC and the computational complexity of Binary MEC are still unknown. It would be interesting to explore whether Property 2, that we used in this paper for achieving a direct 2-approximation algorithm for Binary MEC and a FPT algorithm for Gapless MEC, is also useful for answering to these open questions. Similarly, the design of practical FPT algorithms for the general MEC parameterized by the fragment length is an interesting research direction.

Recent advances in sequencing technologies are radically changing the characteristics of the produced data. For example, long gapless reads with sequencing errors uniformly distributed will likely be common in the near future. The design of FPT algorithms that exploit these characteristics is another important research direction. Furthermore, the drop of sequencing costs allows large-scale studies of rare diseases. In fact, they are usually caused by rare mutations that can only be reliably discovered by sequencing many related individuals. As a consequence, we expect an increasing interest in the study of new formulations extending MEC on structured populations, as well as in a further investigation of those presented in [12].

Acknowledgements. This work has been stimulated by discussions between PB, GK, and NP during the No.045 NII Shonan workshop on Exact Algorithms for Bioinformatics Research, March 2014, Japan.

References

1. Aguiar, D., Istrail, S.: HapCompass: A fast cycle basis algorithm for accurate haplotype assembly of sequence data. J. of Computational Biology 19(6), 577–590 (2012)
2. Bansal, V., Bafna, V.: HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. Bioinformatics 24(16), i153–i159 (2008)
3. Bonizzoni, P., Della Vedova, G., Dondi, R., Li, J.: The haplotyping problem: An overview of computational models and solutions. J. Comput. Sci. Technol. 18(6), 675–688 (2003)
4. Browning, B., Browning, S.: Haplotypic analysis of wellcome trust case control consortium data. Human Genetics 123(3), 273–280 (2008)
5. Chen, Z.Z., Deng, F., Wang, L.: Exact algorithms for haplotype assembly from whole-genome sequence data. Bioinformatics 29(16), 1938–45 (2013)
6. Cilibrasi, R., van Iersel, L., Kelk, S., Tromp, J.: The complexity of the single individual SNP haplotyping problem. Algorithmica 49(1), 13–36 (2007)
7. Duitama, J., Huebsch, T., McEwen, G., Suk, E., Hoehe, M.R.: ReFHap: a reliable and fast algorithm for single individual haplotyping. In: BCB. pp. 160–169. ACM (2010)
8. Fouilhoux, P., Mahjoub, A.: Solving VLSI design and DNA sequencing problems using bipartization of graphs. Computational Optimization and Applications 51(2), 749–781 (2012)
9. Garey, M.R., Johnson, D.S.: Computers and Intractability: A Guide to the Theory of NP-Completeness. W. H. Freeman (1979)
10. Garg, N., Vazirani, V.V., Yannakakis, M.: Approximate max-flow min-(multi) cut theorems and their applications. SIAM Journal on Computing 25(2), 235–251 (1996)
11. Guo, J., Gramm, J., Hüffner, F., Niedermeier, R., Wernicke, S.: Compression-based fixed-parameter algorithms for feedback vertex set and edge bipartization. J. Comput. Syst. Sci. 72(8), 1386–1396 (2006)
12. Halldórsson, B.V., Aguiar, D., Istrail, S.: Haplotype phasing by multi-assembly of shared haplotypes: Phase-dependent interactions between rare variants. In: PSB. pp. 88–99. World Scientific Publishing (2011)
13. He, D., Choi, A., Pipatsrisawat, K., Darwiche, A., Eskin, E.: Optimal algorithms for haplotype assembly from whole-genome sequence data. Bioinformatics 26(12), i183–i190 (2010)
14. Jiao, Y., Xu, J., Li, M.: On the k-closest substring and k-consensus pattern problems. In: CPM. LNCS, vol. 3109, pp. 130–144 (2004)
15. Khot, S.: On the power of unique 2-prover 1-round games. In: STOC. pp. 767–775 (2002)
16. Kleinberg, J.M., Papadimitriou, C.H., Raghavan, P.: Segmentation problems. In: STOC. pp. 473–482 (1998)
17. Lancia, G., Bafna, V., Istrail, S., Lippert, R., Schwartz, R.: SNPs problems, complexity, and algorithms. In: ESA. LNCS, vol. 2161, pp. 182–193 (2001)
18. Lippert, R., Schwartz, R., Lancia, G., Istrail, S.: Algorithmic strategies for the single nucleotide polymorphism haplotype assembly problem. Briefings in bioinformatics 3(1), 23–31 (2002)
19. Ostrovsky, R., Rabani, Y.: Polynomial-time approximation schemes for geometric min-sum median clustering. J. ACM 49(2), 139–156 (2002)
20. Patterson, M., Marschall, T., Pisanti, N., van Iersel, L., Stougie, L., Klau, G.W., Schönhuth, A.: WhatsHap: Haplotype assembly for future-generation sequencing reads. In: RECOMB. LNCS, vol. 8394, pp. 237–249 (2014)
21. Reed, B.A., Smith, K., Vetta, A.: Finding odd cycle transversals. Oper. Res. Lett. 32(4), 299–301 (2004)

Appendix

Proofs of the lemmas

Lemma 8. *If there exists a conflict free matrix $\mathcal{M}''$ obtained from $\mathcal{M}$ on the first $i-1$ rows that induces a tripartition T' for the columns in $\mathcal{A}(i-1)$, and if T is a tripartition for the columns in $\mathcal{A}(i)$ extending T', then there exists a conflict free matrix $\mathcal{M}'$ obtained from $\mathcal{M}$ on the first i rows that induces the tripartition T for the columns in $\mathcal{A}(i)$.*

Proof. By definition, $p_j[i] \neq -$ and $p_j[y] = -$ for each column $p_j \in \mathcal{A}(i) \setminus \mathcal{A}(i-1)$ and for each $y < i$. By assumption T extends T', hence build $\mathcal{M}'$ such that the columns covered by the first $i-1$ rows are tripartitioned as in $\mathcal{M}''$ and the remaining columns only covered by f_i are tripartitioned according to T. By construction, $\mathcal{M}'$ induces the tripartition T for $\mathcal{A}(i)$. Since $\mathcal{M}''$ is conflict free, it follows that $\mathcal{M}'$ is conflict free by Proposition 7.

Lemma 9. *Consider a gapless fragment matrix $\mathcal{M}$.*

1. *If $D[i,T] = h$, then there exists a conflict free matrix $\mathcal{M}'$ obtained from $\mathcal{M}$ on the first i rows with h corrections that induces a tripartition T for the columns in $\mathcal{A}(i)$.*
2. *If $\mathcal{M}'$ is a conflict free matrix obtained from $\mathcal{M}$ on the first i rows with h corrections that induces a tripartition T for the columns in $\mathcal{A}(i)$, then $D[i,T] \leq h$.*

Proof. We prove the lemma by induction on the number n of rows of $\mathcal{M}$. Both the statements obviously hold for $i = 1$. Assume that lemma holds for $i-1$, we show that both the statements hold for i.

(1) By Eq. (1), there exists a tripartition T' for $\mathcal{A}(i-1)$ such that T extends T' and $D[i,T] = h = \Delta(i,T) + D[i-1,T']$. Assuming $D[i-1,T'] = h'$, by induction there exists a conflict free matrix $\mathcal{M}''$ obtained from $\mathcal{M}$ on the first $i-1$ rows with h' corrections that induces a tripartition T' for $\mathcal{A}(i-1)$. By Proposition 8, there exists a conflict free matrix $\mathcal{M}'$ obtained from $\mathcal{M}$ on the first i rows that induces a tripartition T for $\mathcal{A}(i)$. Since T extends T', by construction we can add $\Delta(i,T)$ corrections on fragment f_i in order to build $\mathcal{M}'$ starting from $\mathcal{M}''$. It follows that $\mathcal{M}'$ is obtained from $\mathcal{M}$ with $\Delta(i,T) + h' = h$ corrections.

(2) Assume that $\mathcal{M}''$ is the submatrix of $\mathcal{M}'$ obtained from $\mathcal{M}$ on the first $i-1$ rows with h' corrections that induces a tripartition T' for $\mathcal{A}(i-1)$. Clearly, T' is extended by T due to the fact that $\mathcal{M}''$ is equal to $\mathcal{M}'$ on the first $i-1$ rows. Since $\mathcal{M}'$ contains $\Delta(i,T)$ corrections on the row f_i by construction, it follows that $h = \Delta(i,T) + h'$. Moreover, we know that $D[i-1,T'] \leq h'$ by induction and by Eq. (1) that $D[i,T] = \Delta(i,T) + \min_{T'' \text{ extended by } T} D[i-1,T'']$. Hence, since $\min_{T'' \text{ extended by } T} D[i-1,T''] \leq D[i-1,T']$, we conclude that $D[i,T] \leq \Delta(i,T) + h'$ and, consequently, $D[i,T] \leq h$.