

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

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

DISCo, computer science department of the University of Milano-Bicocca

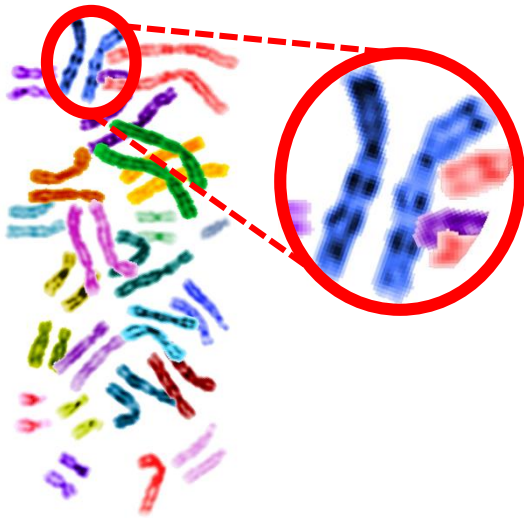
*simone.zaccaria@disco.unimib.it



CPM 2015

26th Annual Symposium on Combinatorial Pattern Matching
Ischia Island, Italy, June 29-July 1, 2015

Haplotypes



Paternal
Haplotype

Maternal
Haplotype

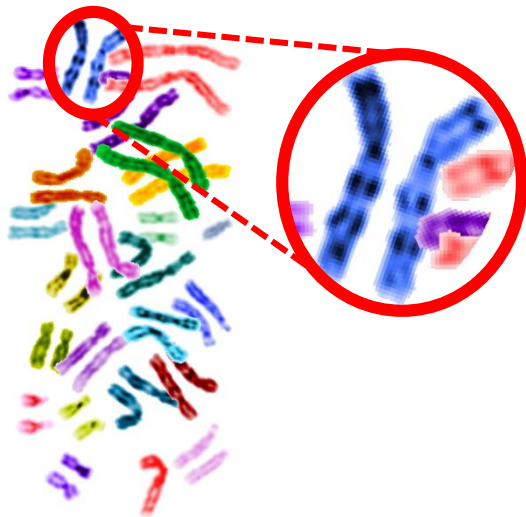
0	- - - - -	0
1	- - - - -	0
0	- - - - -	1
0	- - - - -	1

- 0 major allele
- 1 minor allele

Haplotypes are fundamental in the study of genetic variations

(Browning and Browning, NAR, 2011)

Haplotype Assembly

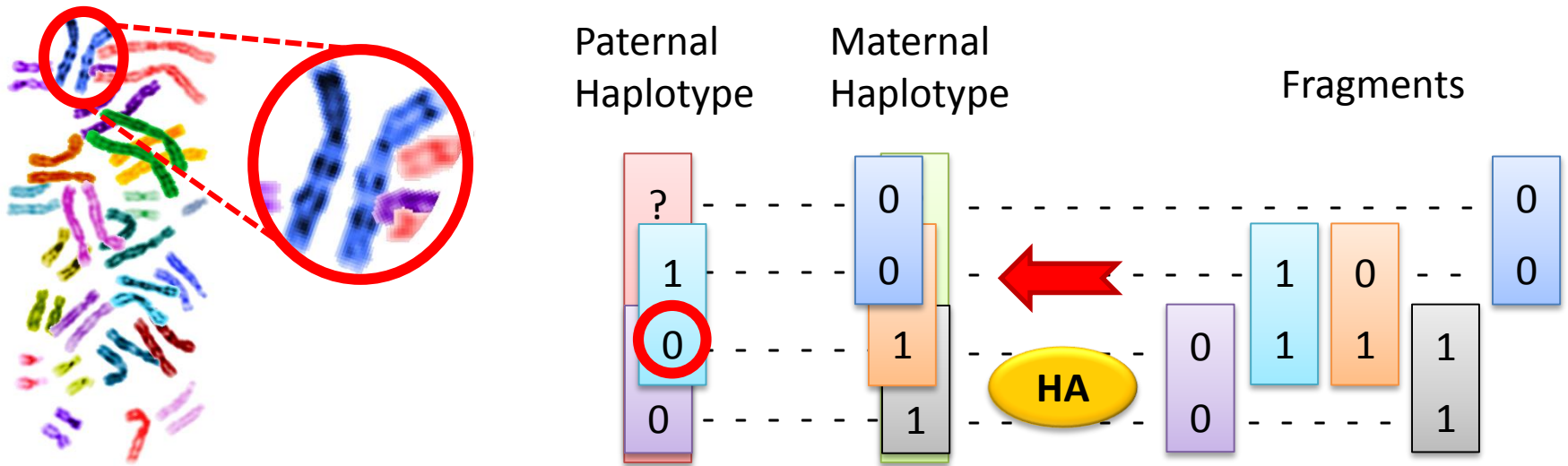


Paternal Haplotype	Maternal Haplotype	Fragments			
0	0	1	0	1	0
1	0				
0	1	0	1	1	0
0	1				

Haplotypes are fundamental in the study of genetic variations
(Browning and Browning, NAR, 2011)

- **Haplotype:** difficult and expensive to obtain
- **Aligned Fragments:** Easy and cost-effective to obtain

Minimum Error Correction



- (sequencing/mapping) errors lead to an optimization problem
- Several formulations have been proposed in literature (Lippert et al., Brief. In Bioinformatics, 2002): Minimum Error Correction (MEC) is one of the most prominent

Fragment Matrix

← Genome positions →

← fragments →	1	0	-	1	1	1	-	0	1	-
	-	-	0	0	1	-	1	0	1	-
	-	1	0	0	0	-	-	1	-	1
	-	-	0	1	0	-	1	1	-	0
	-	-	-	0	0	0	1	1	1	1

- Fragment matrix, on n fragments and m SNP positions:
 - Each row corresponds to a fragment
 - Each column correspond to a genome position

Interesting Parameters

9

1	0	-	1	1	1	-	0	1	-
-	-	0	0	1	-	1	0	1	-
-	1	0	0	0	-	-	1	-	1
-	-	0	1	0	-	1	1	-	0
-	-	-	0	0	0	1	1	1	1

- Interesting and (practical) parameters:
 - Fragment length ℓ

Interesting Parameters

1	0	-	1	1	1	-	0	1	-
-	-	0	0	1	-	1	0	1	-
-	1	0	0	0	-	-	1	-	1
-	-	0	1	0	-	1	1	-	0
-	-	-	0	0	0	1	1	1	1

3

- Interesting and (practical) parameters:
 - Fragment length ℓ
 - **Coverage**

Minimum Error Correction (MEC) problem

← Genome positions →

	1	0	-	1	1	1	-	0	1	-
← fragments →	-	-	0	0	1	-	1	0	1	-
	-	1	0	0	0	-	-	1	-	1
	-	-	0	1	0	-	1	1	-	0
	-	-	-	0	0	0	1	1	1	1

- **Input:** Fragment matrix
- **Output:** Minimum number of error corrections that allow to bipartition the fragments without internal conflict (**conflict free**)

Minimum Error Correction (MEC) problem

← Genome positions →

← fragments →

1	0	-	1	1	1	-	0	1	-
-	-	0	0	1	-	1	0	1	-
-	1	0	0	0	-	-	1	-	1
-	-	0	1	0	-	1	1	-	0
-	-	-	0	0	0	1	1	1	1

Conflicts

- **Input:** Fragment matrix
- **Output:** Minimum number of error corrections that allow to bipartition the fragments without internal conflict (**conflict free**)

Minimum Error Correction (MEC) problem



- **Input:** Fragment matrix
- **Output:** Minimum number of error corrections that allow to bipartition the fragments without internal conflict (**conflict free**)

Variants: Holes and Gaps

(Cilibrasi et al., Algorithmica, 2007)

MEC
(holes, gaps)

0	0	-	-	1	1	-	-	-	-
-	0	-	-	0	-	0	-	-	-
-	1	1	-	-	1	0	-	-	-

Gapless MEC
(holes, no gaps)

0	0	1	0	1	1	-	-	-	-
-	0	0	1	0	1	0	-	-	-
-	-	-	-	1	0	1	0	1	0

Binary MEC
(no holes, no gaps)

0	0	1	0	1	1	0	1	0	1
0	0	0	1	0	1	0	1	0	1
0	0	1	1	1	0	1	0	1	0

Variant Motivations

(Cilibrasi et al., Algorithmica, 2007)

MEC (holes, gaps)

- Represent any instance for the Haplotype Assembly
 - Fit for both paired-end and single-end reads
-

Gapless MEC (holes, no gaps)

- Represent single-end reads without gaps
-

Binary MEC (no holes, no gaps)

- Unrealistic for the Haplotype Assembly
- Interesting from a mathematical point of view: variant of the well-known Hamming 2-Median Clustering Problem

Contributions

*(Cilibrasi et al., Algorithmica, 2007)

	NP	APX	FPT by ℓ	Other FPT
MEC (holes, gaps)	NP-hard *	APX-hard *	With restrictive assumptions (He et al., Bioinformatics, 2013)	FPT by coverage (Patterson et al., RECOMB, 2014)
Gapless MEC (holes, no gaps)	NP-hard *	?		
Binary MEC (no holes, no gaps)	?	PTAS *		

Contributions of **this paper**

	NP	APX	FPT by ℓ	Other FPT
MEC (holes, gaps)	NP-hard	APX-hard not in APX log-apx	With restrictive assumptions (He et al., Bioinformatics, 2013)	• FPT by coverage (Patterson et al., RECOMB, 2014) • FPT by total number of corrections
Gapless MEC (holes, no gaps)	NP-hard	?	Without any assumption	
Binary MEC (no holes, no gaps)	?	PTAS (More direct) 2-apx algo.		

Heterozygous and Homozygous

← SNP positions →

← fragments →	1	0	-	1	1	1	-	0	1	-
	-	1	0	0	0	-	-	1	-	1
	-	-	0	0	1	-	1	0	1	-
	-	-	0	1	0	-	1	1	-	0
	-	-	-	0	0	0	1	1	1	1

- Columns = vectors belonging to $\{0, 1, -\}^n$:
 - **Heterozygous** columns contain both 0's and 1's
=> encode a bipartition of the covered fragments
 - **Homozygous** columns belong to $\{0, -\}^n$ or $\{1, -\}^n$
=> no information on the bipartition

Heterozygous Columns

← SNP positions →

← fragments →	1	0	-	1	1	1	-	0	1	-
	-	1	0	0	0	-	-	1	-	1
	-	-	0	0	1	-	1	0	1	-
	-	-	0	1	0	-	1	1	-	0
	-	-	-	0	0	0	1	1	1	1

- Columns = vectors belonging to $\{0, 1, -\}^n$:
 - Heterozygous columns contain both 0's and 1's**
=> **encode a bipartition of the covered fragments**
 - Homozygous columns belong to $\{0, -\}^n$ or $\{1, -\}^n$
=> no information on the bipartition

Heterozygous Columns

← SNP positions →

← fragments →

1	0	-	1	1	1	-	0	1	-
-	1	0	0	0	-	-	1	-	1
-	-	0	0	1	-	1	0	1	-
-	-	0	1	0	-	1	1	-	0
-	-	-	0	0	0	1	1	1	1

- Columns = vectors belonging to $\{0, 1, -\}^n$:
 - Heterozygous columns contain both 0's and 1's**
 => **encode a bipartition of the covered fragments**
 - Homozygous columns belong to $\{0, -\}^n$ or $\{1, -\}^n$
 => no information on the bipartition

Heterozygous and Homozygous

← SNP positions →

← fragments →	1	0	-	1	1	1	-	0	1	-
	-	1	0	0	0	-	-	1	-	1
	-	-	0	0	1	-	1	0	1	-
	-	-	0	1	0	-	1	1	-	0
	-	-	-	0	0	0	1	1	1	1

- Columns = vectors belonging to $\{0, 1, -\}^n$:
 - Heterozygous columns contain both 0's and 1's
=> encode a bipartition of the covered fragments
 - **Homozygous columns belong to $\{0, -\}^n$ or $\{1, -\}^n$**
=> no information on the bipartition

Accordance

1	0
0	0
0	0
1	0
0	0
-	0

1	-
0	0
0	0
1	1
0	0
-	1

1	-
0	1
0	1
1	0
0	1
-	0

- Two columns are in accordance if and only if:
 - At least one is homozygous
 - They are all equal/complementary on common fragments

Accordance

1	0
0	0
0	0
1	0
0	0
-	0

1	-
0	0
0	0
1	1
0	0
-	1

1	-
0	1
0	1
1	0
0	1
-	0

- Two columns are in accordance if and only if:
 - At least one is homozygous
 - They are all equal/complementary on common fragments

Accordance

1	0
0	0
0	0
1	0
0	0
-	0

1	-
0	0
0	0
1	1
0	0
-	1

1	-
0	1
0	1
1	0
0	1
-	0

- Two columns are in accordance if and only if:
 - **At least one is homozygous**
 - They are all equal/complementary on common fragments

Accordance

1	0
0	0
0	0
1	0
0	0
-	0

1	-
0	0
0	0
1	1
0	0
-	1

1	-
0	1
0	1
1	0
0	1
-	0

- Two columns are in accordance if and only if:
 - At least one is homozygous
 - **They are all equal/complementary on common fragments**

Lemma of Accordance

0	1	-
-	0	1
1	1	0

A **gapless** fragment matrix is conflict free

if and only if

each pair of columns is in accordance

Lemma 1: counter-example

0	1	-
-	0	1
1	-	0

A ~~gapless~~ fragment matrix is conflict free

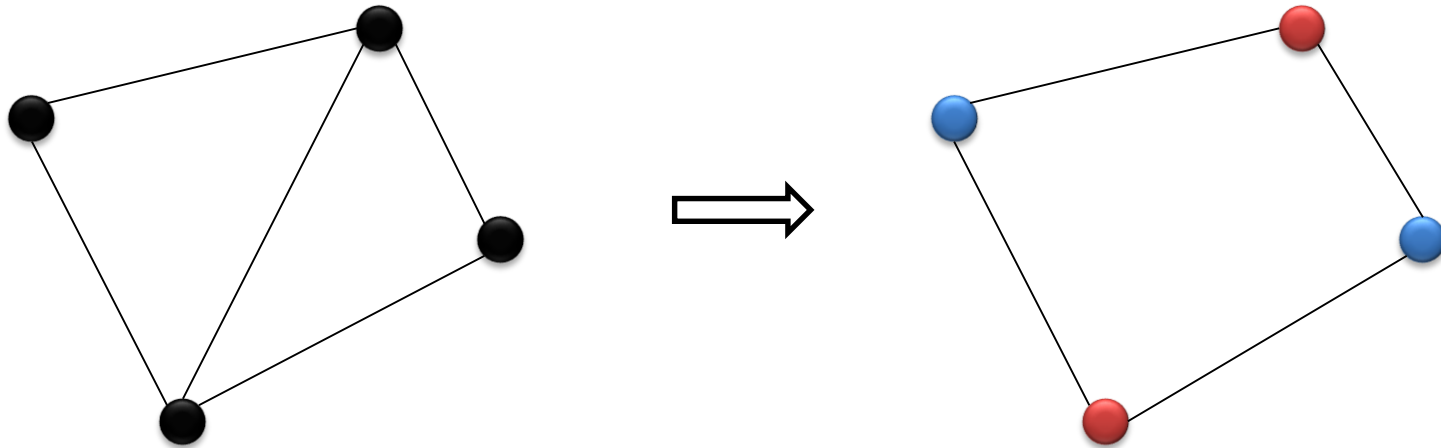
~~it and only if~~

each pair of columns is in accordance

Results

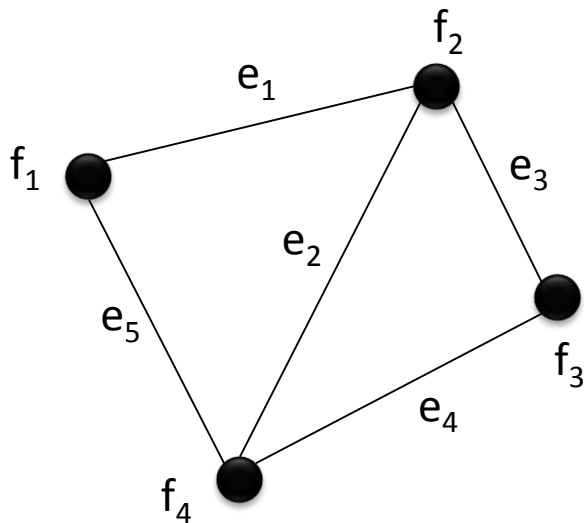
- Contributions:
 1. **Inapproximability of MEC (plus approximation and FPT results)**
 2. Gapless MEC is in FPT when parameterized by the fragment length
 3. A 2-approximation algorithm for Binary MEC

Edge Bipartization (EB)



- **Input:** an undirected graph $G = (V, E)$
- **Output:** $E' \subseteq E$ of minimum size such that $G' = (V, E \setminus E')$ is bipartite

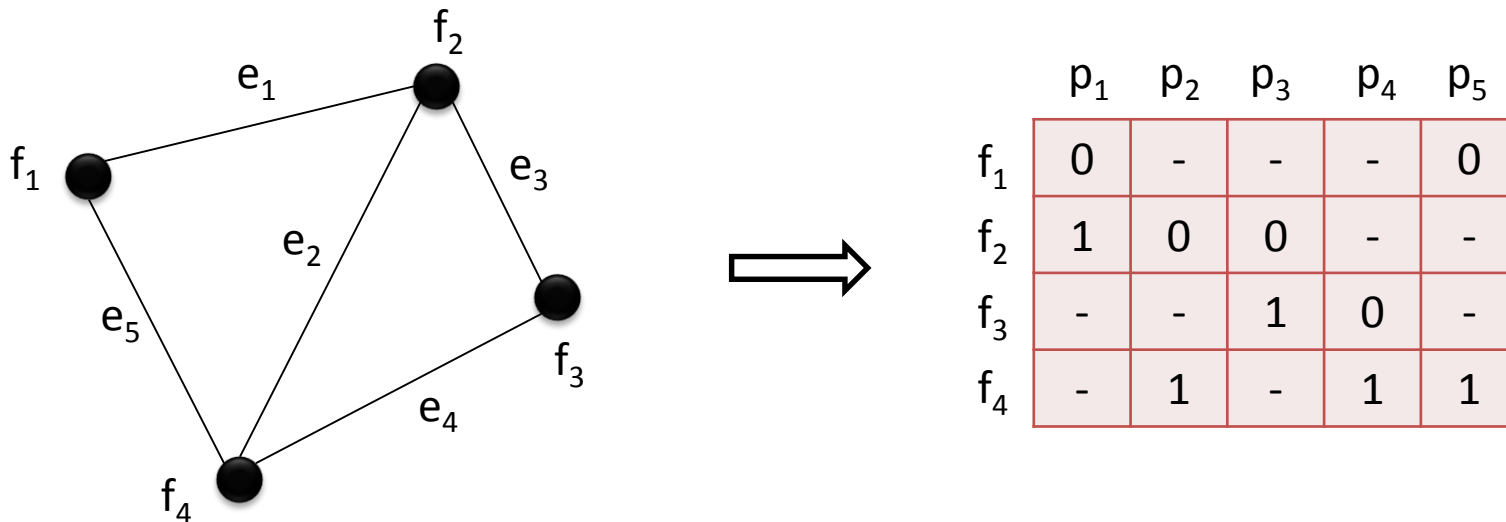
From EB to MEC



	p_1	p_2	p_3	p_4	p_5
f_1	0	-	-	-	0
f_2	1	0	0	-	-
f_3	-	-	1	0	-
f_4	-	1	-	1	1

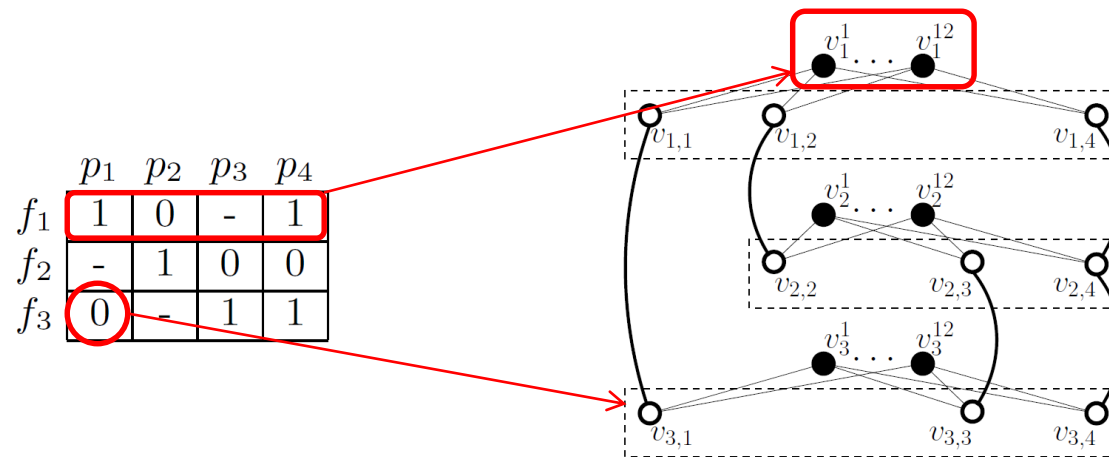
- The resulting fragment matrix has: $|V|$ rows and $|E|$ columns
- row = vertex / column = edge assigning 0/1 to the extremities
- At most 1 correction per column $\Rightarrow$ correction = removal

Inapproximability of MEC



- Khot proved that, under the Unique Games Conjecture, EB is not in APX
=> Under the Unique Games Conjecture, **MEC is not in APX**

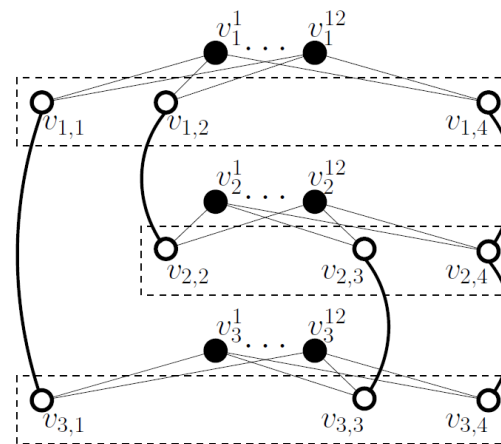
Novel Fragment Graph



- Variation of the reduction used by (Fouilhoux *et al.*, 2012) to reduce MEC to the Maximum Induced Bipartite Subgraph problem (MIBS)
- Novel version of the fragment graph:
 - Entry-nodes
 - Fragment-nodes (with $mn+1$ duplicates)

From MEC to GB

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



- Results from reducing MEC to the well-known Graph Bipartization (GB):
 - MEC is in FPT when parameterized by total number of corrections
 - MEC can be approximated in polynomial time within factor $O(\log nm)$

Results

- Contributions:
 1. Inapproximability of MEC (plus approximation and FPT results)
 2. **Gapless MEC is in FPT when parameterized by the fragment length**
 3. A 2-approximation algorithm for Binary MEC

Lemma 2

0	1	0	0	-	-
1	0	0	1	0	-
-	1	0	0	-	-
-	-	0	0	1	-
-	-	0	1	0	1

A gapless fragment matrix is conflict free

If and only if

Columns can be tripartited (L, H, R) such that H are homozygous, L/R are complementary and R/R or L/L are equal

Lemma 2

0	1	0	0	-	-
1	0	0	1	0	-
-	1	0	0	-	-
-	-	0	0	1	-
-	-	0	1	0	1

M is conflict free

If and only if

Columns can be tripartited (L, H, R) such that H are homozygous, L/R are complementary and R/R or L/L are equal

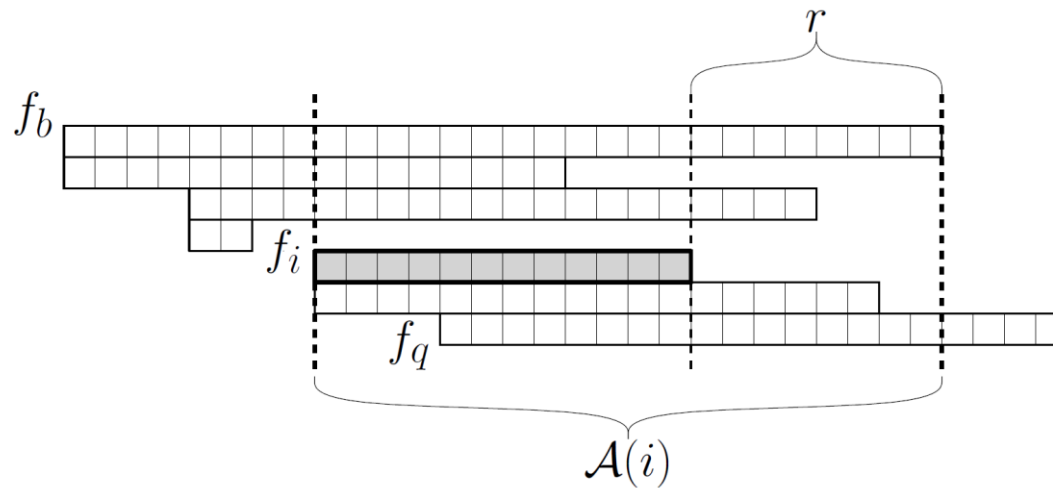
Main Idea

- Goal: build the tripartition in an optimal way
- 2 basic ideas:
 1. Any tripartition can be built row-wise adding a subset of columns at each step
 2. The subset of columns (active columns) has a size less or equal to maximum fragment length ℓ

Main Idea

- Goal: build the tripartition in an optimal way
- 2 basic ideas:
 1. Any tripartition can be built row-wise adding a subset of columns at each step
 2. The subset of columns (active columns) has a size less or equal to maximum fragment length ℓ

Active Columns



- Assume that reads are sorted by starting positions
- Active columns for f_i : covered by f_i +
covered by both previous and next fragments
- The size is less than ℓ

Dynamic Programming

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

- The algorithm is based on a dynamic programming approach

Recursive Step

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

- Search for the best T' for $A(i - 1)$ extended by T for $A(i - 1)$
 - Extending = equivalent partitioning
-
- The algorithm is based on a dynamic programming approach

Local Contribution

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

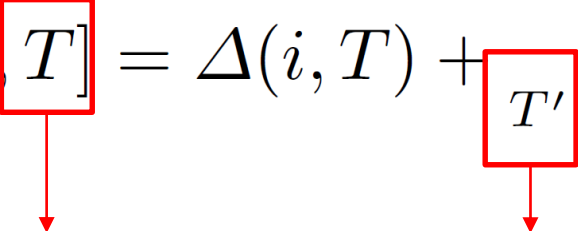
The minimum number of corrections on f_i to build T :

- Transform columns in H into homozygous
- Choose the best combination 0/1 for L/R

- The algorithm is based on a dynamic programming approach:

Time Complexity

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



$O(3^\ell)$ $O(3^\ell)$

- The algorithm is based on a dynamic programming approach:
- Time complexity: $O(3^{2\ell} \cdot \ell \cdot n)$ ($O(3^\ell \cdot \ell \cdot n)$ storing partial info.)

Results

- Contributions:
 1. Inapproximability of MEC (plus approximation and FPT results)
 2. Gapless MEC is in FPT when parameterized by the fragment length
 3. **A 2-approximation algorithm for Binary MEC**

Basic Observation

1	1	0	0	1	0
1	0	0	1	0	1
0	1	1	1	1	0
0	1	0	0	1	0
1	0	0	0	0	0

- **Notice:** In this variant, each column encodes a feasible solution, that is a bipartition for all the fragments

Iteration

p_0

1	1	0	0	1	0
1	0	0	1	0	1
0	1	1	1	1	0
0	1	0	0	1	0
1	0	0	0	0	0

- Iteratively choose a column and assume that it is the closest to the optimal bipartition

Build Accordance

p_o

0	1	0	0	1	0
1	0	0	1	0	1
0	1	1	1	1	0
0	1	0	0	1	0
1	0	0	0	0	0

- Assume that p_o is the closest column to optimal bipartition
- Correct each other column with minimum number of corrections in accordance to p_o :
 - Correct into column equal/complementary to p_o
 - Correct into homozygous column

Build Accordance

p_o

0	1	0	0	1	0
1	0	0	1	0	1
0	1	1	1	1	0
0	1	0	0	1	0
1	0	0	0	0	0

- Assume that p_o is the closest column to optimal bipartition
- Correct each other column with minimum number of corrections in accordance to p_o :
 - Correct into column equal/complementary to p_o
 - Correct into homozygous column

Build Accordance

p_o

0	1	0	0	1	0
1	0	0	1	0	1
0	1	0	1	1	0
0	1	0	0	1	0
1	0	0	0	0	0

- Assume that p_o is the closest column to optimal bipartition
- Correct each other column with minimum number of corrections in accordance to p_o :
 - Correct into column equal/complementary to p_o
 - Correct into homozygous column

Build Accordance

p_0

0	1	0	0	1	0
1	0	0	1	0	1
0	1	0	0	1	0
0	1	0	0	1	0
1	0	0	1	0	0

- Assume that p_0 is the closest column to optimal bipartition
- Correct each other column with minimum number of corrections in accordance to p_0 :
 - Correct into column equal/complementary to p_0
 - Correct into homozygous column

Build Accordance

p_o

0	1	0	0	1	0
1	0	0	1	0	0
0	1	0	0	1	0
0	1	0	0	1	0
1	0	0	1	0	0

- Assume that p_o is the closest column to optimal bipartition
- Correct each other column with minimum number of corrections in accordance to p_o :
 - Correct into column equal/complementary to p_o
 - Correct into homozygous column

2-approximation

p_0

0	1	0	0	1	0
1	0	0	1	0	0
0	1	0	0	1	0
0	1	0	0	1	0
1	0	0	1	0	0

- Using the triangle inequality in a proper way, the algorithm can be easily proved to be a 2-approximation
- Return the solution with minimum cost OPT' in $O(m^2n)$ time

Future Directions

- Gaps can be modelled as zero-weight elements
=> FPT result for gapless variant holds even for general MEC
- Many NGS data from genomes with many chromosomes
(plants, fishes, tumors. ...)
=> Fixed Parameter Tractability of k-ploid MEC by using novel combinatorial properties, here introduced?
- Novel variants of MEC in order to:
 - exploit the novel characteristics of future-generation sequencing technologies (such as uniform distribution of sequencing errors)
 - extend haplotype assembly on structured populations (such as trios)

Thanks for the attention!
Questions?

Correspondence to:

simone.zaccaria@disco.unimib.it

Personal page:

<http://algotlab.eu/simone-zaccaria>