Improving an Exact Solution to the (l,d) Planted Motif Problem

Maria Clara Isabel Sia

October 12, 2015

Introduction

- The (l,d) planted motif problem
- Definitions of key concepts

EMS-GT

- The EMS-GT algorithm
- Efficiency strategies

Methodology

- Research objectives
- Development and evaluation

Results

- Pattern-based speedup technique
- Performance

Conclusion

Introduction

Introduction

- motifs: repeated sub-sequences in DNA that have some biological significance

Introduction

- motifs: repeated sub-sequences in DNA that have some biological significance

- DNA motif finding: search for motifs over a set of DNA sequences, allowing for mismatches due to mutation

Introduction

- motifs: repeated sub-sequences in DNA that have some biological significance

- DNA motif finding: search for motifs over a set of DNA sequences, allowing for mismatches due to mutation

- known as a difficult problem in computational biology and CS (proven NP-complete)

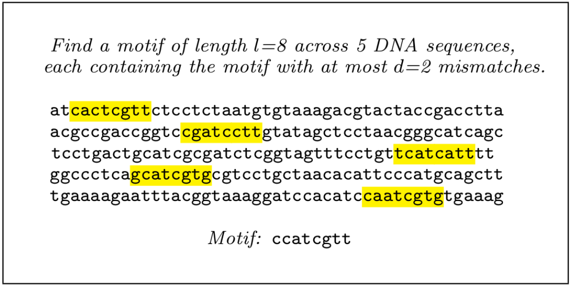
The (*l*,*d*) planted motif problem

Find a motif of length l=8 across 5 DNA sequences, each containing the motif with at most d=2 mismatches.

```
atcactcgttctcctctaatgtgtaaagacgtactaccgacctta
acgccgaccggtccgatccttgtatagctcctaacgggcatcagc
tcctgactgcatcgcgatctcggtagtttcctgttcatcatttt
ggccctcagcatcgtgcgtcctgctaacacattcccatgcagctt
tgaaaagaatttacggtaaaggatccacatccaatcgtgtgaaag
```

Motif: ccatcgtt

The (l,d) planted motif problem



Find a motif of length l=8 across 5 DNA sequences, each containing the motif with at most d=2 mismatches.

at**cactcgtt**ctcctctaatgtgtaaagacgtactaccgacctta
acgccgaccggtc**cgatcctt**gtatagctcctaacgggcatcagc
tcctgactgcatcgcgatctcggtagtttcctgt**tcatcatt**tt
ggccctca**gcatcgtg**cgtcctgctaacacattcccatgcagctt
tgaaaagaatttacggtaaaggatccacatc**caatcgtg**tgaaag

Motif: ccatcgtt

Given a set $\mathcal{S} = \{S_1, ...S_n\}$ of n DNA sequences of length L, find M, the set of sub-sequences (motifs) of length $l < L$ which occur with at most d mismatches in each sequence in $\mathcal{S}$.

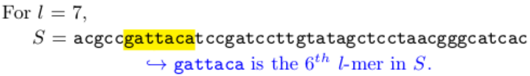
Definitions of key concepts

- l-mer
- Hamming distance $dH(x_1, x_2)$
- d-neighborhood $N(x, d)$ of l-mer x
- d-neighborhood $\mathcal{N}(S, d)$ of sequence S

Definitions of key concepts

- *l*-mer

  - sequence of length l



For $l = 7$,

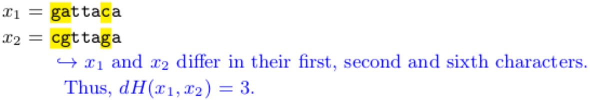
$S =$ acgccgattacatccgatccttgtatagctcctaacgggcatcac

↪ gattaca is the 6^{th} l-mer in S.

- Hamming distance $dH(x_1, x_2)$
- d-neighborhood $N(x, d)$ of l-mer x
- d-neighborhood $\mathcal{N}(S, d)$ of sequence S

Definitions of key concepts

- l-mer
- Hamming distance $dH(x_1, x_2)$
  - number of mismatches between l-mers x_1 and x_2

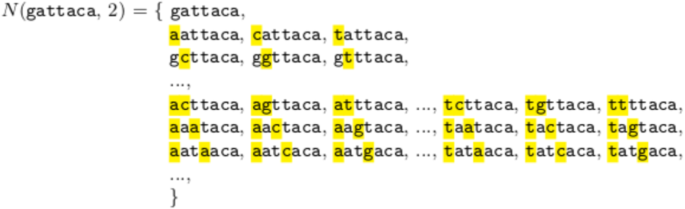


$x_1 =$ gattaca
$x_2 =$ cgttaga
↪ x_1 and x_2 differ in their first, second and sixth characters.
Thus, $dH(x_1, x_2) = 3$.

- d-neighborhood $N(x, d)$ of l-mer x
- d-neighborhood $\mathcal{N}(S, d)$ of sequence S

Definitions of key concepts

- *l*-mer
- Hamming distance $dH(x_1, x_2)$
- d-neighborhood $N(x, d)$ of l-mer x
  - set of all l-mers having at most d mismatches with x



N(gattaca, 2) = { gattaca,
aattaca, cattaca, tattaca,
gcttaca, ggttaca, gtttaca,
...,
acttaca, agttaca, atttaca, ..., tcttaca, tgttaca, ttttaca,
aaataca, aactaca, aagtaca, ..., taataca, tactaca, tagtaca,
aataaca, aatcaca, aatgaca, ..., tataaca, tatcaca, tatgaca,
...,
}

- d-neighborhood $\mathcal{N}(S, d)$ of sequence S

Definitions of key concepts

- *l*-mer
- Hamming distance $dH(x_1, x_2)$
- *d*-neighborhood $N(x, d)$ of *l*-mer x
- *d*-neighborhood $\mathcal{N}(S, d)$ of sequence S
  - union of *d*-neighborhoods of all *l*-mers in S

$$S = \texttt{acgccgattacatccgatccttgtatagctcctaacgggcatcac}$$

$$\mathcal{N}(S, 2) = N(\texttt{acgccga}, 2) \cup N(\texttt{cgccgat}, 2) \cup ... \cup N(\texttt{gcatcac}, 2)$$

$\hookrightarrow$ d-neighborhood of first l-mer in S

Definitions of key concepts

- l-mer
- Hamming distance $dH(x_1, x_2)$
- d-neighborhood $N(x, d)$ of l-mer x
- d-neighborhood $\mathcal{N}(S, d)$ of sequence S

Introduction

The (l,d) planted motif problem

Definitions of key concepts

EMS-GT

The EMS-GT algorithm

Efficiency strategies

Methodology

Research objectives

Development and evaluation

Results

Pattern-based speedup technique

Performance

Conclusion

The EMS-GT algorithm

The EMS-GT algorithm proceeds in two main steps:

The EMS-GT algorithm

The EMS-GT algorithm proceeds in two main steps:

1. *Generate candidates*

The EMS-GT algorithm

The EMS-GT algorithm proceeds in two main steps:

1. *Generate candidates*
   Take the intersection of the d-neighborhoods of the first n' sequences $S_1, S_2, ..., S_{n'}$. Every l-mer in the resulting set C is a candidate motif.

The EMS-GT algorithm

The EMS-GT algorithm proceeds in two main steps:

1. *Generate candidates*
   Take the intersection of the d-neighborhoods of the first n' sequences $S_1, S_2, ..., S_{n'}$. Every l-mer in the resulting set C is a candidate motif.
2. *Test candidates*

The EMS-GT algorithm

The EMS-GT algorithm proceeds in two main steps:

1. *Generate candidates*
   Take the intersection of the d-neighborhoods of the first n' sequences $S_1, S_2, ..., S_{n'}$. Every l-mer in the resulting set C is a candidate motif.

2. *Test candidates*
   For every candidate c in C, check whether a d-neighbor of c appears in each of the remaining sequences $S_{n'+1}, S_{n'+2}, ...S_n$. If this is the case, accept c as a motif.

EMS-GT efficiency strategies

- *l*-mer enumeration
- Bit-based set representation
- Recursive neighborhood generation

EMS-GT efficiency strategies

- *l*-mer enumeration
- Bit-based set representation
- Recursive neighborhood generation

EMS-GT efficiency strategies

- *l*-mer enumeration
- Bit-based set representation
- Recursive neighborhood generation

EMS-GT efficiency strategies

- *l*-mer enumeration
- Bit-based set representation
- Recursive neighborhood generation

Introduction

- The (l,d) planted motif problem
- Definitions of key concepts

EMS-GT

- The EMS-GT algorithm
- Efficiency strategies

Methodology

- Research objectives
- Development and evaluation

Results

- Pattern-based speedup technique
- Performance

Conclusion

Methodology

The main objectives of this research are:

Methodology

The main objectives of this research are:

1. To develop a speedup technique for EMS-GT that takes advantage of distance-related patterns in the search space;

Methodology

The main objectives of this research are:

1. To develop a speedup technique for EMS-GT that takes advantage of distance-related patterns in the search space;
2. To evaluate the speedup technique with regard to improvement in runtime; and

Methodology

The main objectives of this research are:

1. To develop a speedup technique for EMS-GT that takes advantage of distance-related patterns in the search space;
2. To evaluate the speedup technique with regard to improvement in runtime; and
3. To evaluate the improved version of EMS-GT against state-of-the-art motif search algorithms.

Methodology

- *Developing an EMS-GT speedup technique*

Methodology

- *Developing an EMS-GT speedup technique*
  In EMS-GT, l-mer neighborhoods are represented with repeating patterns of bits. We exploit these patterns in a speedup technique that sets bits quickly and in blocks.

Methodology

- *Developing an EMS-GT speedup technique*
  In EMS-GT, l-mer neighborhoods are represented with repeating patterns of bits. We exploit these patterns in a speedup technique that sets bits quickly and in blocks.

- *Evaluating performance*

Methodology

- *Developing an EMS-GT speedup technique*
  In EMS-GT, l-mer neighborhoods are represented with repeating patterns of bits. We exploit these patterns in a speedup technique that sets bits quickly and in blocks.

- *Evaluating performance*
  Performance is benchmarked on "challenging" (l,d) instances: (9,2), (11,3), (13,4), (15,5) and (17,6).

Methodology

- *Developing an EMS-GT speedup technique*
  In EMS-GT, l-mer neighborhoods are represented with repeating patterns of bits. We exploit these patterns in a speedup technique that sets bits quickly and in blocks.

- *Evaluating performance*
  Performance is benchmarked on "challenging" (l,d) instances: (9,2), (11,3), (13,4), (15,5) and (17,6).

  - *Synthetic datasets*

Methodology

- *Developing an EMS-GT speedup technique*
  In EMS-GT, l-mer neighborhoods are represented with repeating patterns of bits. We exploit these patterns in a speedup technique that sets bits quickly and in blocks.

- *Evaluating performance*
  Performance is benchmarked on "challenging" (l,d) instances: (9,2), (11,3), (13,4), (15,5) and (17,6).

  - *Synthetic datasets*
    - sets of 20 randomly-generated DNA sequences of length 600, with the same (l,d) motif planted once in each sequence.

Pattern-based speedup technique

Performance improvement with speedup

Performance vs PMS8, qPMS9

Introduction

The (l,d) planted motif problem

Definitions of key concepts

EMS-GT

The EMS-GT algorithm

Efficiency strategies

Methodology

Research objectives

Development and evaluation

Results

Pattern-based speedup technique

Performance

Conclusion

Thank you!