A simple Motif finder based on random projections

Presented by

Gergana Stanilova

Department of Mathematics and Computer Science

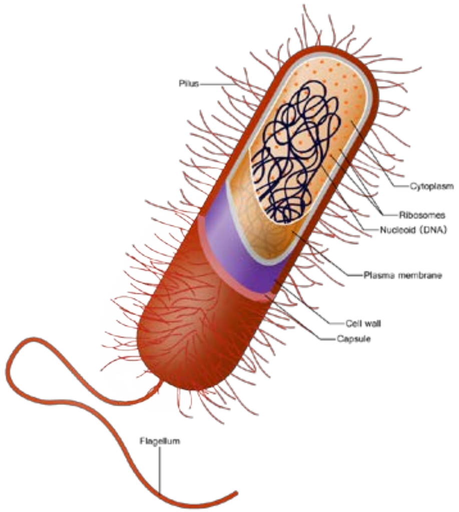
Freie Universität Berlin

Supervisor: Christopher Pockrandt

Outline

- ➢ Introduction & Motivation
- Problem
- Background
- Solution
- Validation
- Future work

Introduction & Motivation



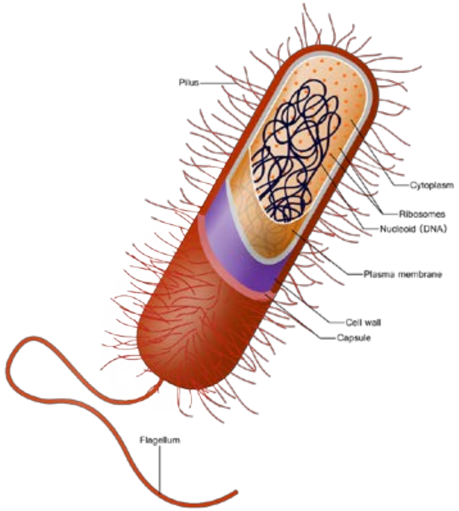
*

Bacteria

* Source: https://upload.wikimedia.org/wikipedia/commons/thumb/c/c5/Prokaryote_cell.svg/914px-Prokaryote_cell.svg.png

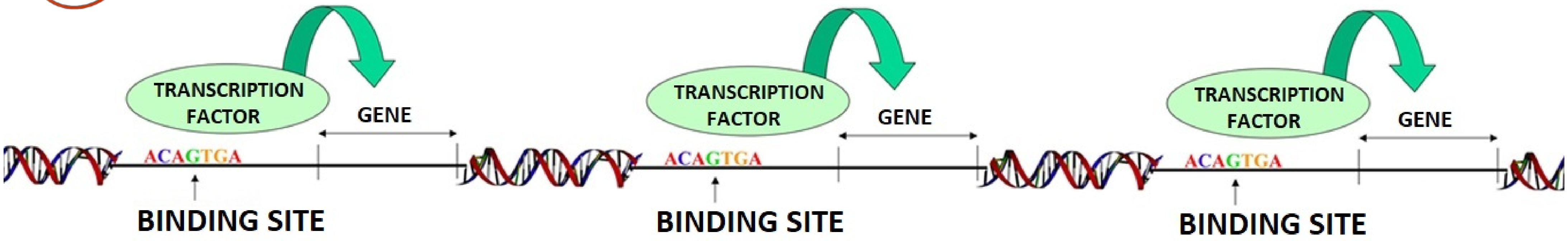
** Source: http://2013.igem.org/wiki/images/c/c3/Xmusoftware_promoter4.jpg

Introduction & Motivation



*

Bacterial genome

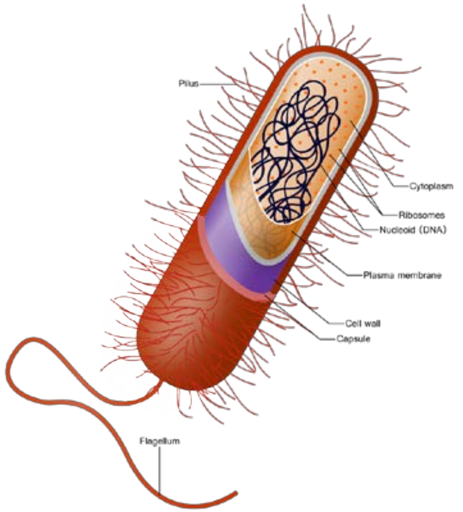


**

* Source: https://upload.wikimedia.org/wikipedia/commons/thumb/c/c5/Prokaryote_cell.svg/914px-Prokaryote_cell.svg.png

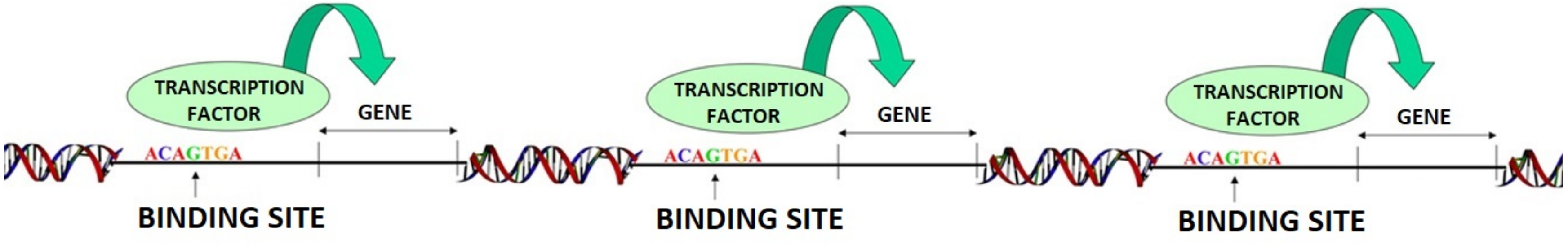
** Source: http://2013.igem.org/wiki/images/c/c3/Xmusoftware_promoter4.jpg

Introduction & Motivation



*

Bacterial genome



**

- Influence the expression of a gene family
- Know the binding sites
- Find a pattern („motif“)

* Source: https://upload.wikimedia.org/wikipedia/commons/thumb/c/c5/Prokaryote_cell.svg/914px-Prokaryote_cell.svg.png

** Source: http://2013.igem.org/wiki/images/c/c3/Xmusoftware_promoter4.jpg

Outline

Introduction & Motivation

➢ Problem

Background

Solution

Validation

Future work

Problem

The (l, d) - Motif problem

Given:

- *t* sequences (the regions upstream of the genes)
- each sequence of length *n*

Wanted

- The motif M
- of length *l*
- with *d* point substitutions (mutations)

1 GGTCTATCTGATTCCAGTCGTCTAT

2 CAATTCCAGACGTCTAAAGGTCTA

3 ACCTTATTCCAGTCGGCTTTCTCTC

4 AGCTAAGAGTCTGATACCAGTCGT

. …

. …

. …

t GGTTTCCAATCGTCTATCCCTGAG

Problem

The (l, d) - Motif problem

Given:

- *t* sequences (the regions upstream of the genes)
- each sequence of length *n*

Wanted

- The motif M
- of length *l*
- with *d* point substitutions (mutations)

1	GGTCTATCTGATTCCAGTCGTCTAT
2	CAATTCCAGACGTCTAAAGGTCTA
3	ACCTTATTCCAGTCGGCTTTCTCTC
4	AGCTAAGAGTCTGATACCAGTCGT
.	…
.	…
.	…
t	GGTTTCCAATCGTCTATCCCTGAG

length *n*

Problem

The (l, d) - Motif problem

Given:

- *t* sequences (the regions upstream of the genes)
- each sequence of length *n*

Wanted

- The motif M
- of length *l*
- with *d* point substitutions (mutations)

1	GGTCTATCTGATTCCAGTCGTCTAT
2	CAATTCCAGACGTCTAAAGGTCTA
3	ACCTTATTCCAGTCGGCTTTCTCTC
4	AGCTAAGAGTCTGATACCAGTCGT
.	…
.	…
.	…
t	GGTTTCCAATCGTCTATCCCTGAG

length *n*

Problem

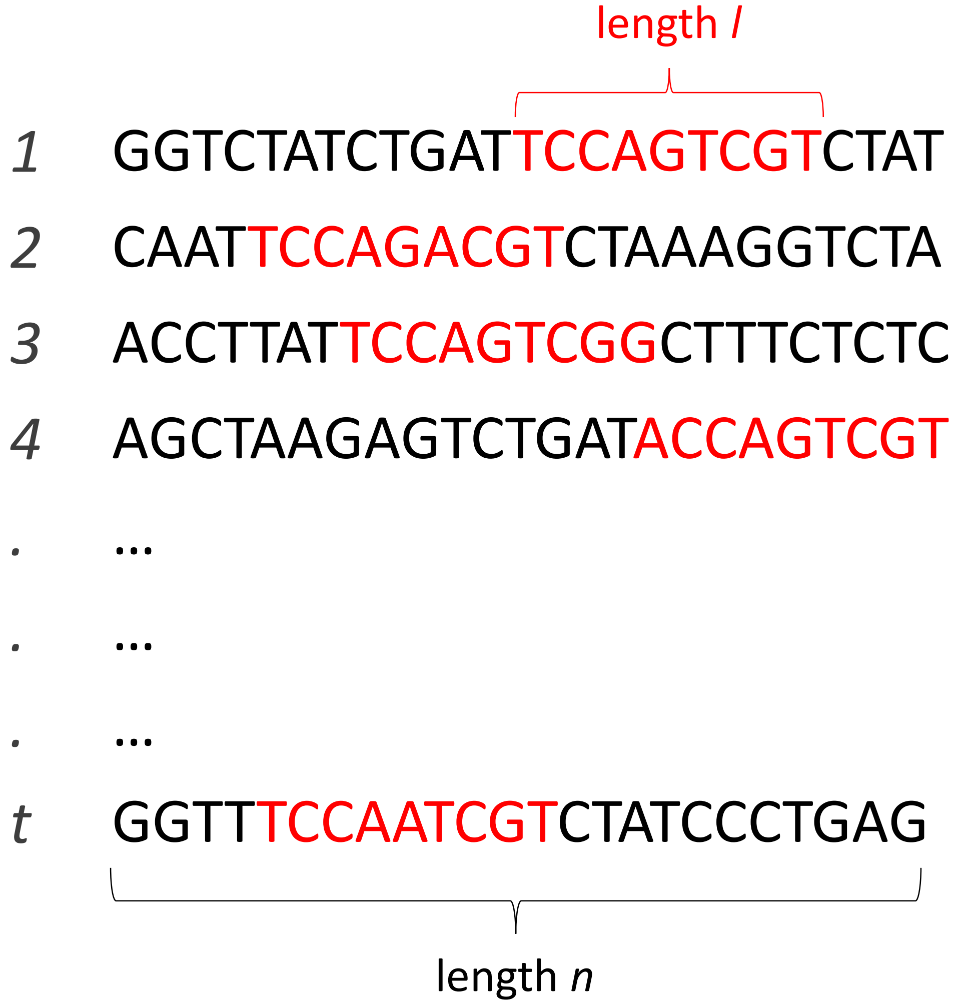
The (l, d) - Motif problem

<u>Given:</u>

- *t* sequences (the regions upstream of the genes)
- each sequence of length *n*

<u>Wanted</u>

- The motif M
- of length *l*
- with *d* point substitutions (mutations)



length *l*

1 GGTCTATCTGATTCCAGTCGTCTAT

2 CAATTCCAGACGTCTAAAGGTCTA

3 ACCTTATTCCAGTCGGCTTTCTCTC

4 AGCTAAGAGTCTGATACCAGTCGT

. …

. …

. …

t GGTTTCCAATCGTCTATCCCTGAG

length *n*

Problem

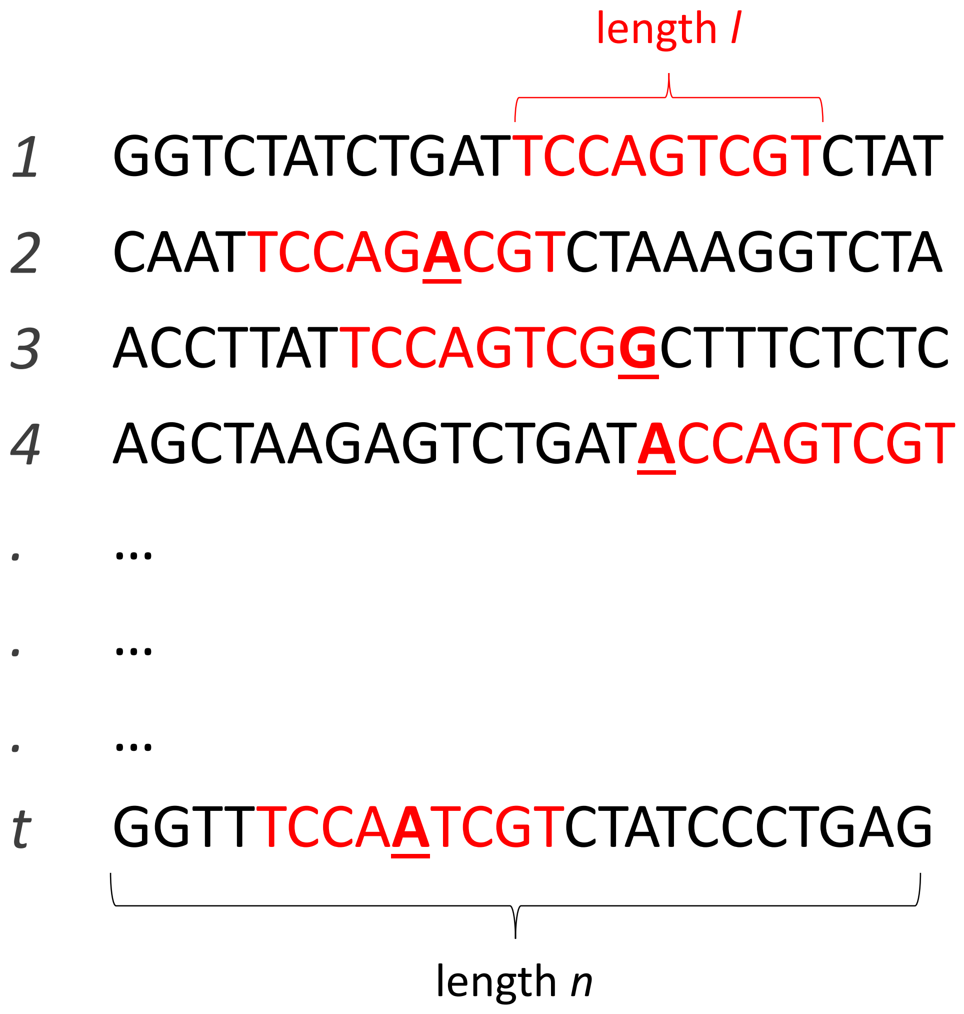
The (l, d) - Motif problem

Given:

- *t* sequences (the regions upstream of the genes)
- each sequence of length *n*

Wanted

- The motif M
- of length *l*
- with *d* point substitutions (mutations)



length *l*

1 GGTCTATCTGATTCCAGTCGTCTAT

2 CAATTCCAG**A**CGTCTAAAGGTCTA

3 ACCTTATTCCAGTCG**G**CTTTCTCTC

4 AGCTAAGAGTCTGAT**A**CCAGTCGT

. …

. …

. …

t GGTTTCCA**A**TCGTCTATCCCTGAG

length *n*

Outline

Introduction & Motivation

Problem

➢ Background

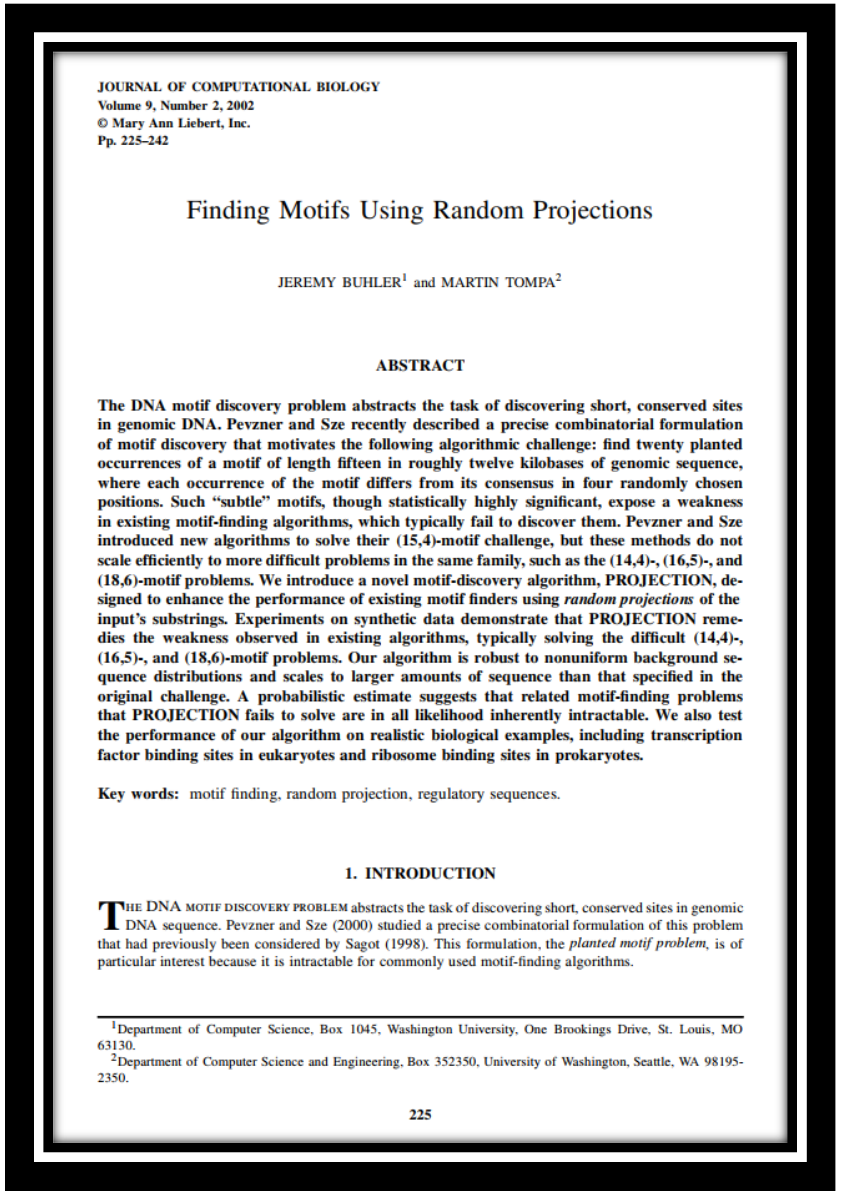
Solution

Validation

Future work

Background

- Gibbs sampling and MEME had a poor performance for the (15,4)-motif problem in terms of accuracy
- Algorithms by Pevzner and Sze fail for (14,4)-, (16,5)-, and (18,6)-motif problems
- PROJECTION



JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 9, Number 2, 2002
© Mary Ann Liebert, Inc.
Pp. 225–242

Finding Motifs Using Random Projections

JEREMY BUHLER[1] and MARTIN TOMPA[2]

ABSTRACT

The DNA motif discovery problem abstracts the task of discovering short, conserved sites in genomic DNA. Pevzner and Sze recently described a precise combinatorial formulation of motif discovery that motivates the following algorithmic challenge: find twenty planted occurrences of a motif of length fifteen in roughly twelve kilobases of genomic sequence, where each occurrence of the motif differs from its consensus in four randomly chosen positions. Such "subtle" motifs, though statistically highly significant, expose a weakness in existing motif-finding algorithms, which typically fail to discover them. Pevzner and Sze introduced new algorithms to solve their (15,4)-motif challenge, but these methods do not scale efficiently to more difficult problems in the same family, such as the (14,4)-, (16,5)-, and (18,6)-motif problems. We introduce a novel motif-discovery algorithm, PROJECTION, designed to enhance the performance of existing motif finders using *random projections* of the input's substrings. Experiments on synthetic data demonstrate that PROJECTION remedies the weakness observed in existing algorithms, typically solving the difficult (14,4)-, (16,5)-, and (18,6)-motif problems. Our algorithm is robust to nonuniform background sequence distributions and scales to larger amounts of sequence than that specified in the original challenge. A probabilistic estimate suggests that related motif-finding problems that PROJECTION fails to solve are in all likelihood inherently intractable. We also test the performance of our algorithm on realistic biological examples, including transcription factor binding sites in eukaryotes and ribosome binding sites in prokaryotes.

Key words: motif finding, random projection, regulatory sequences.

1. INTRODUCTION

THE DNA MOTIF DISCOVERY PROBLEM abstracts the task of discovering short, conserved sites in genomic DNA sequence. Pevzner and Sze (2000) studied a precise combinatorial formulation of this problem that had previously been considered by Sagot (1998). This formulation, the *planted motif problem*, is of particular interest because it is intractable for commonly used motif-finding algorithms.

[1] Department of Computer Science, Box 1045, Washington University, One Brookings Drive, St. Louis, MO 63130.

[2] Department of Computer Science and Engineering, Box 352350, University of Washington, Seattle, WA 98195-2350.

* C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262:208–214, 8 October 1993.
T. L. Bailey and C. Elkan. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. Machine Learning, 21(1-2):51–80, Oct. 1995.
** P. Pevzner and S.-H. Sze. Combinatorial approaches to finding subtle signals in DNA sequences. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 269–278, Aug. 2000.
*** Jeremy Buhler and Martin Tompa. Finding motifs using random projections. J Comput Biol. 2002. 9(2):225-42

Outline

Introduction & Motivation

Problem

Background

➢ Solution

Validation

Future work

Solution

Implementation

- C++
- SEQAN Library *

Approach & Methods

For m trials

- Step 1: Random Projections

- Step 2: Refinement

- Step 3: Consensus sequence

Take the best consensus sequence from all trials

* K. Reinert, T. H. Dadi, M. Ehrhardt, H. Hauswedell, S. Mehringer, R. Rahn, J. Kim, C. Pockrandt, J. Winkler, E. Siragusa, G. Urgese, and D. Weese. The seqan c++ template library for efficient sequence analysis: a resource for programmers. Journal of biotechnology, vol. 261, pp. 157-168, 2017

Solution

Implementation

- C++
- SEQAN Library

Approach & Methods

For m trials

- Step 1: **Random Projections**

- Step 2: Refinement

- Step 3: Consensus sequence

Take the best consensus sequence from all trials

Random Projections

- Project l-mers onto k-mers

GCCACGT *l-mer* *l=7*

Solution

Implementation

- C++
- SEQAN Library

Approach & Methods

For m trials

- Step 1: **Random Projections**

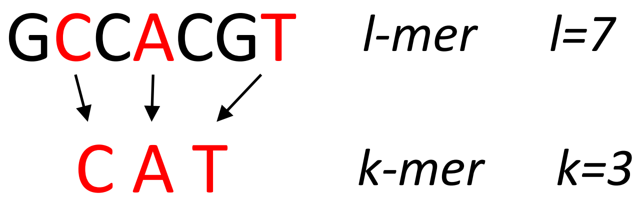
- Step 2: Refinement

- Step 3: Consensus sequence

Take the best consensus sequence from all trials

Random Projections

- Project l-mers onto k-mers



Solution

Implementation

- C++
- SEQAN Library

Approach & Methods

For m trials

- Step 1: **Random Projections**

- Step 2: Refinement

- Step 3: Consensus sequence

Take the best consensus sequence from all trials

Random Projections

- Project l-mers onto k-mers
- Hash the k-mers

hash(C A T) = hashValue

Solution

Implementation

- C++
- SEQAN Library

Approach & Methods

For m trials

- Step 1: **Random Projections**

- Step 2: Refinement

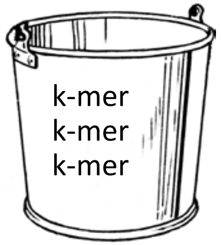
- Step 3: Consensus sequence

Take the best consensus sequence from all trials

Random Projections

- Project l-mers onto k-mers
- Hash the k-mers
- Order them into buckets

hashValue



k-mer
k-mer
k-mer

*

* Source: http://www.marysrosaries.com/collaboration/images/6/68/Pail_or_Bucket_001.png

Solution

Implementation

- C++
- SEQAN Library

Approach & Methods

For m trials

- Step 1: **Random Projections**

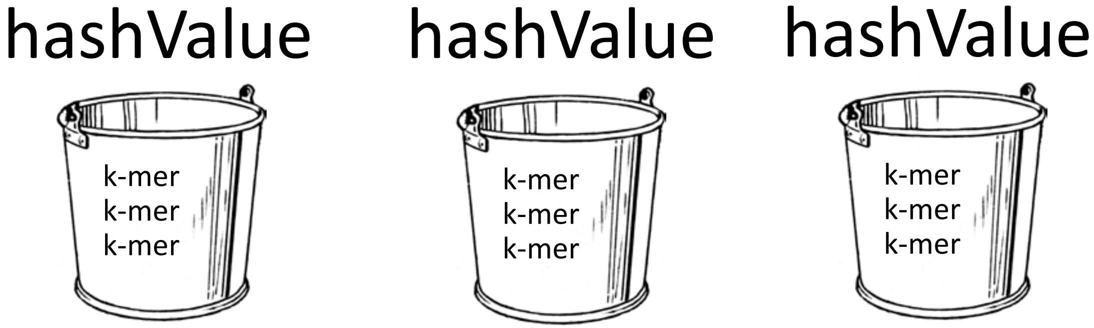
- Step 2: Refinement

- Step 3: Consensus sequence

Take the best consensus sequence from all trials

Random Projections

- Project l-mers onto k-mers
- Hash the k-mers
- Order them into buckets



Solution

Implementation

- C++
- SEQAN Library

Approach & Methods

For m trials

- Step 1: Random Projections

- Step 2: **Refinement**

- Step 3: Consensus sequence

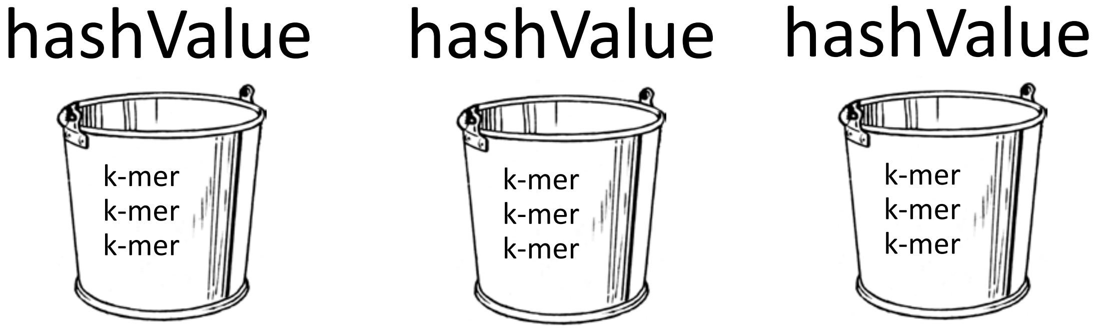
Take the best consensus sequence from all trials

Refinement

EM-Algorithm:

- known: the sequences

- unknown: the positions at which the motif occurs



Solution

Implementation

- C++
- SEQAN Library

Approach & Methods

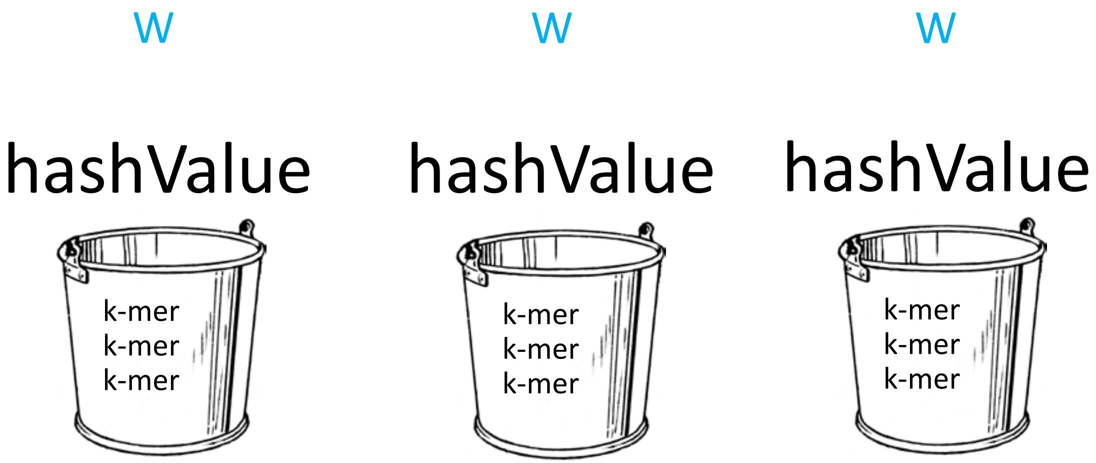
For m trials

- Step 1: Random Projections
- Step 2: **Refinement**
- Step 3: Consensus sequence

Take the best consensus sequence from all trials

Refinement

- For each bucket with at least *s* elements create a weight matrix: with what frequency does each base occur?



W W W

hashValue hashValue hashValue

k-mer
k-mer
k-mer

k-mer
k-mer
k-mer

k-mer
k-mer
k-mer

Solution

Implementation

- C++
- SEQAN Library

Approach & Methods

For m trials

- Step 1: Random Projections

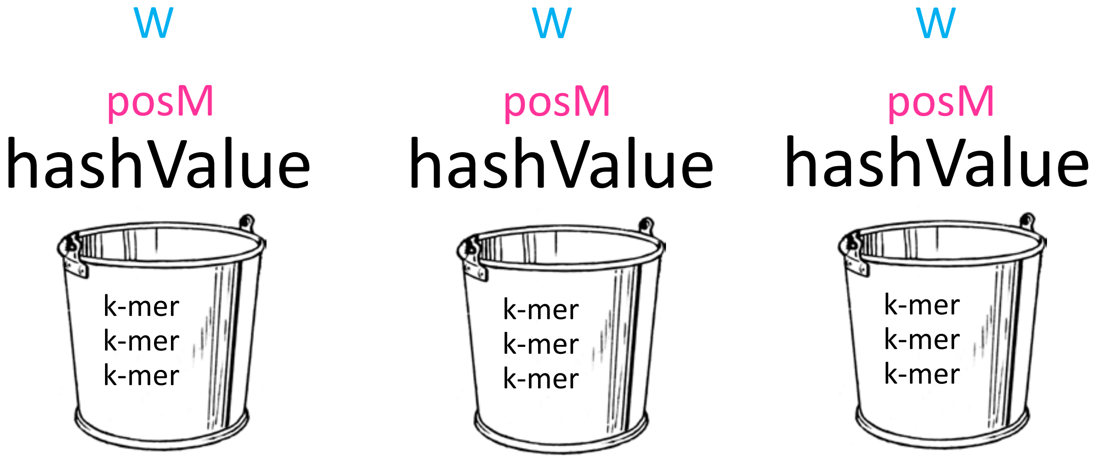
- Step 2: **Refinement**

- Step 3: Consensus sequence

Take the best consensus sequence from all trials

Refinement

- For each bucket create a weight matrix
- Create a position matrix: what is the most probable start position of the motif in each sequence?



W W W

posM posM posM

hashValue hashValue hashValue

k-mer
k-mer
k-mer

k-mer
k-mer
k-mer

k-mer
k-mer
k-mer

Solution

Implementation

- C++
- SEQAN Library

Approach & Methods

For m trials

- Step 1: Random Projections

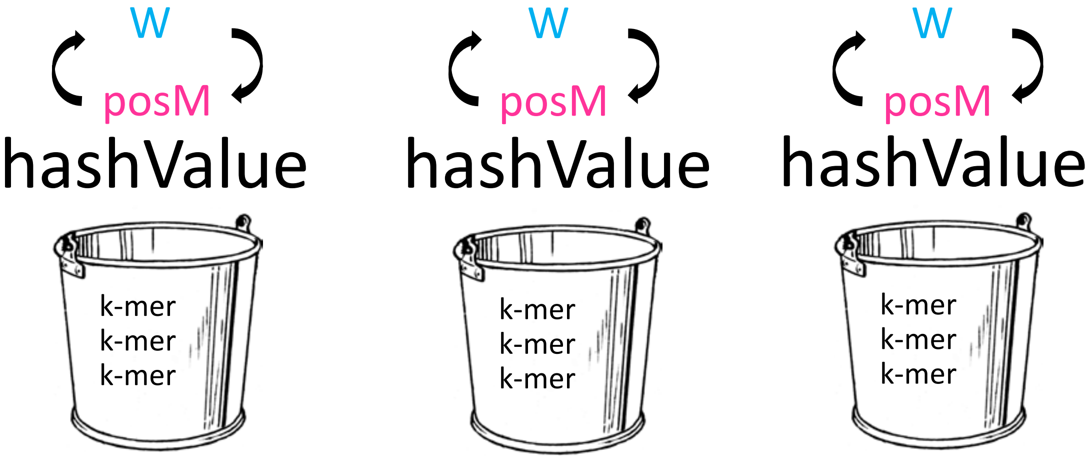
- Step 2: **Refinement**

- Step 3: Consensus sequence

Take the best consensus sequence from all trials

Refinement

- For each bucket create a weight matrix
- Create a position matrix
- Refine the weight matrix until convergence



Solution

Implementation

- C++
- SEQAN Library

Approach & Methods

For m trials

- Step 1: Random Projections

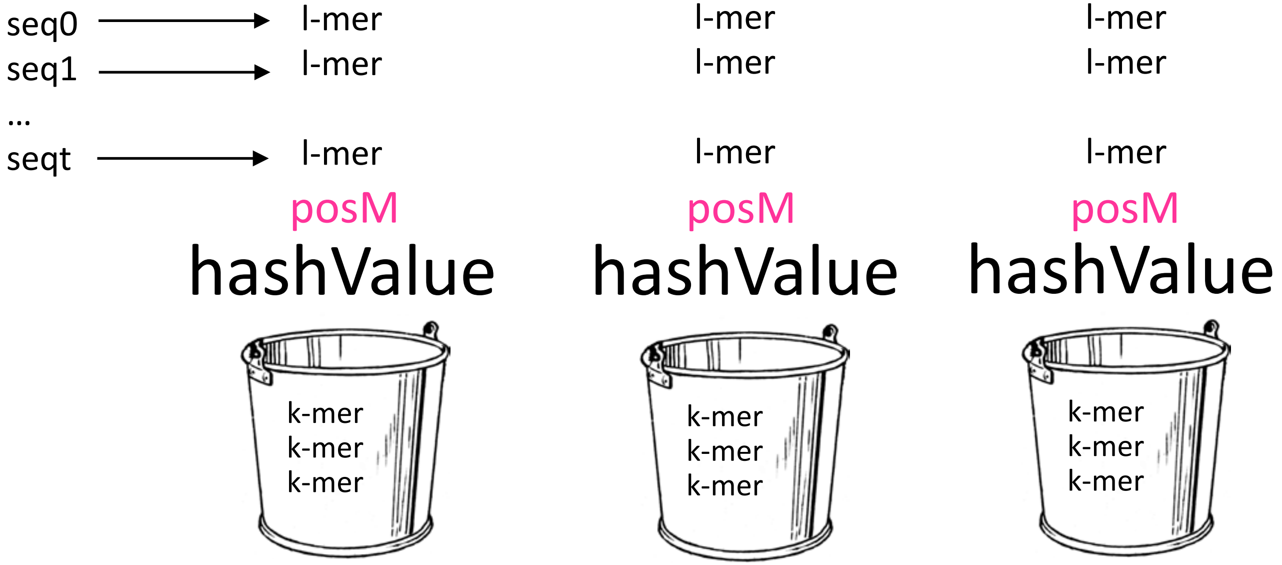
- Step 2: Refinement

- Step 3: **Consensus sequence**

Take the best consensus sequence from all trials

Consensus sequence

- Extract an l-mer from each sequence using the position matrix



Solution

Implementation

- C++
- SEQAN Library

Approach & Methods

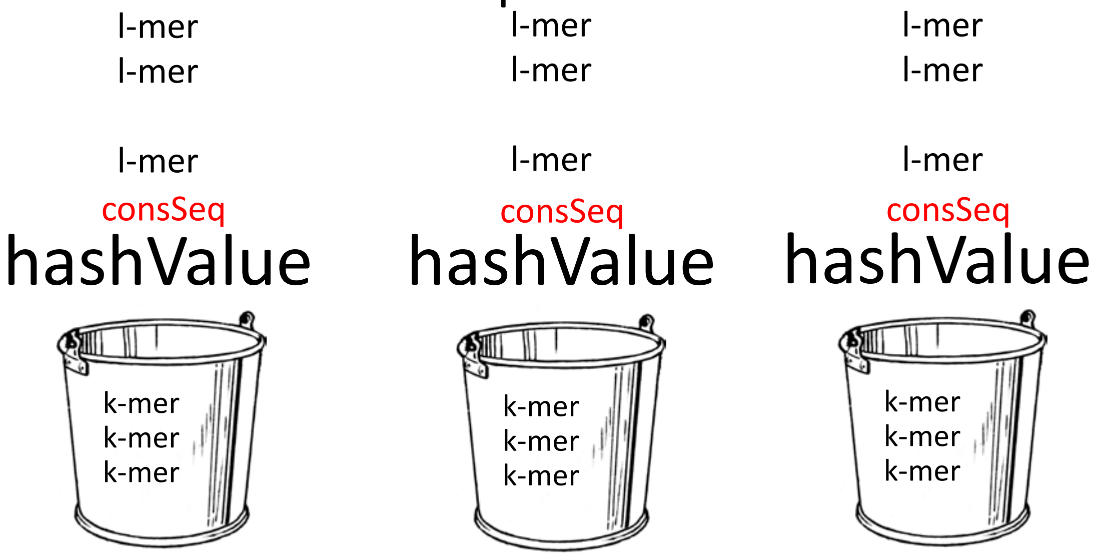
For m trials

- Step 1: Random Projections
- Step 2: Refinement
- Step 3: **Consensus sequence**

Take the best consensus sequence from all trials

Consensus sequence

- Extract an l-mer from each sequence
- Create a consensus sequence



l-mer
l-mer

l-mer
consSeq
hashValue
k-mer
k-mer
k-mer

l-mer
l-mer

l-mer
consSeq
hashValue
k-mer
k-mer
k-mer

l-mer
l-mer

l-mer
consSeq
hashValue
k-mer
k-mer
k-mer

Solution

Implementation

- C++
- SEQAN Library

Approach & Methods

For m trials

- Step 1: Random Projections

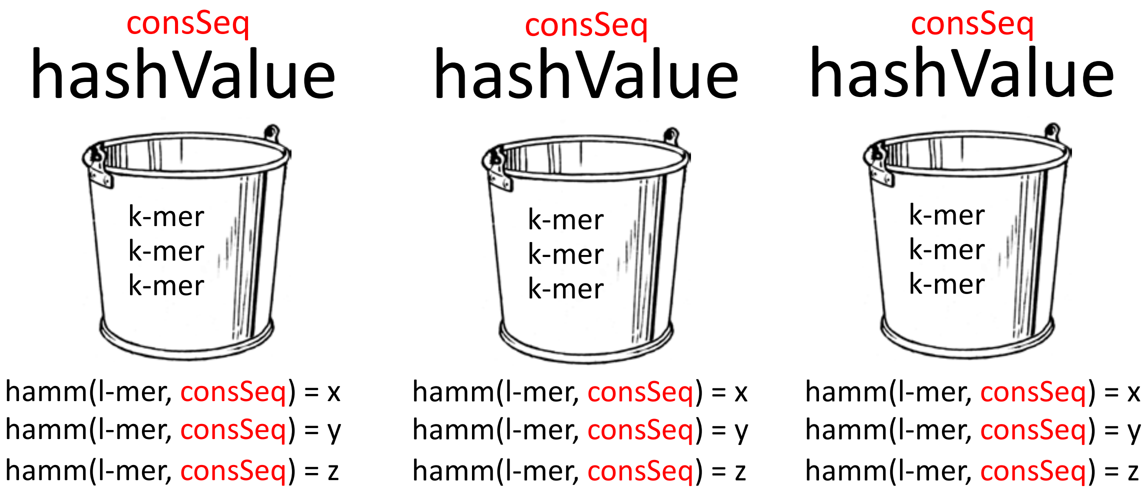
- Step 2: Refinement

- Step 3: **Consensus sequence**

Take the best consensus sequence from all trials

Consensus sequence

- Extract an l-mer from each sequence
- Create a consensus sequence
- Calculate the Hamming distance between each l-mer and the consensus sequence



consSeq
hashValue
k-mer
k-mer
k-mer
hamm(l-mer, consSeq) = x
hamm(l-mer, consSeq) = y
hamm(l-mer, consSeq) = z

consSeq
hashValue
k-mer
k-mer
k-mer
hamm(l-mer, consSeq) = x
hamm(l-mer, consSeq) = y
hamm(l-mer, consSeq) = z

consSeq
hashValue
k-mer
k-mer
k-mer
hamm(l-mer, consSeq) = x
hamm(l-mer, consSeq) = y
hamm(l-mer, consSeq) = z

Solution

Implementation

- C++
- SEQAN Library

Approach & Methods

For m trials

- Step 1: Random Projections

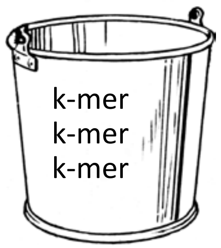
- Step 2: Refinement

- Step 3: **Consensus sequence**

Take the best consensus sequence from all trials

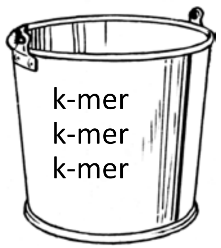
Consensus sequence

- Extract an l-mer from each sequence
- Create a consensus sequence
- Calculate the Hamming distance
- Calculate a score for each bucket: the number of hamming distances < *d* (the max number of mutations)



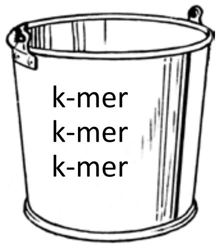
k-mer
k-mer
k-mer

consSeq => score



k-mer
k-mer
k-mer

consSeq => score



k-mer
k-mer
k-mer

consSeq => score

Solution

Implementation

- C++
- SEQAN Library

Approach & Methods

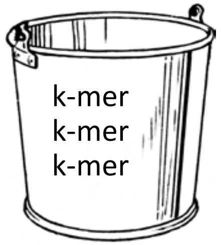
For m trials

- Step 1: Random Projections
- Step 2: Refinement
- Step 3: **Consensus sequence**

Take the best consensus sequence from all trials

Consensus sequence

- Extract an l-mer from each sequence
- Create a consensus sequence
- Calculate the Hamming distance
- Calculate a score for each bucket
- Keep the consensus sequence from the bucket with the lowest score



consSeq => score

Solution

Implementation

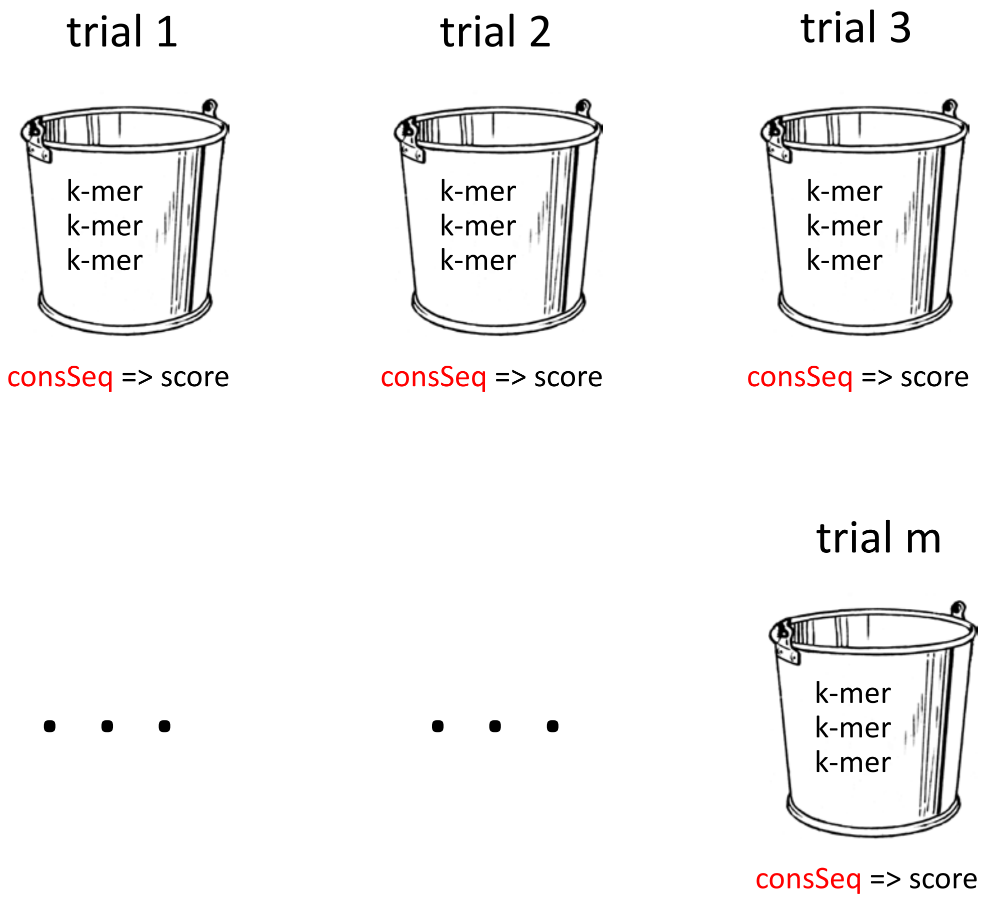
- C++
- SEQAN Library

Approach & Methods

For m trials

- Step 1: Random Projections
- Step 2: Refinement
- Step 3: Consensus sequence

Take the best consensus sequence from all trials



Solution

Implementation

- C++
- SEQAN Library

Approach & Methods

For m trials

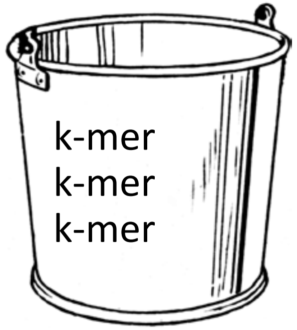
- Step 1: Random Projections

- Step 2: Refinement

- Step 3: Consensus sequence

Take the best consensus sequence from all trials

trial x



The consensus sequence should be the planted motif

Example

Planted motif: AGCTC

- Length of motif *l = 5*
- Maximum number of mutations *d = 2*

Example

Planted motif: AGCTC

- Length of motif *l = 5*
- Maximum number of mutations *d = 2*

Given sequences:

CCGCGAGCTC

AGATCGTAAC

TGGGCTACCT

- Number of sequences *t = 3*
- Length of one sequence *n = 10*

Example

Planted motif: AGCTC

- Length of motif $l = 5$
- Maximum number of mutations $d = 2$

Given sequences:

CCGCG**AGCTC**

AGATCGTAAC

TG**GGCTA**CCT

- Number of sequences $t = 3$
- Length of one sequence $n = 10$

Example

Planted motif: AGCTC

- Length of motif *l = 5*
- Maximum number of mutations *d = 2*

Given sequences:

CCGCG<u>**AGCTC**</u>	*0 mutations*
<u>**AGATC**</u>GTAAC	*1 mutation*
TG<u>**GGCTA**</u>CCT	*2 mutations*

- Number of sequences *t = 3*
- Length of one sequence *n = 10*

Example

Planted motif: AGCTC

- Length of motif *l = 5*
- Maximum number of mutations *d = 2*

Given sequences:

CCGCGAGCTC	*0 mutations*
AGATCGTAAC	*1 mutation*
TGGGCTACCT	*2 mutations*

- Number of sequences *t = 3*
- Length of one sequence *n = 10*

Step 1: Random projections

- From an l-mer to a k-mer
- Choice of k?

Example

Planted motif: AGCTC

- Length of motif $l = 5$
- Maximum number of mutations $d = 2$

Given sequences:

CCGCGAGCTC	*0 mutations*
AGATCGTAAC	*1 mutation*
TGGGCTACCT	*2 mutations*

- Number of sequences $t = 3$
- Length of one sequence $n = 10$

Step 1: Random projections

- From an l-mer to a k-mer
- Choice of k? -> **k < l - d**

Example

Planted motif: AGCTC

- Length of motif *l = 5*
- Maximum number of mutations *d = 2*

Given sequences:

CCGCGAGCTC *0 mutations*

AGATCGTAAC *1 mutation*

TGGGCTACCT *2 mutations*

- Number of sequences *t = 3*
- Length of one sequence *n = 10*

Step 1: Random projections

- From an l-mer to a k-mer
- Choice of k? -> **k < 5 - 2**

Example

Planted motif: AGCTC

- Length of motif *l = 5*
- Maximum number of mutations *d = 2*

Given sequences:

CCGCGAGCTC	*0 mutations*
AGATCGTAAC	*1 mutation*
TGGGCTACCT	*2 mutations*

- Number of sequences *t = 3*
- Length of one sequence *n = 10*

Step 1: Random projections

- From an l-mer to a k-mer
- Choice of k? -> **k < 3** -> **k = 2**

Example

Planted motif: AGCTC

- Length of motif *l = 5*
- Maximum number of mutations *d = 2*

Given sequences:

CCGCGAGCTC *0 mutations*

AGATCGTAAC *1 mutation*

TGGGCTACCT *2 mutations*

- Number of sequences *t = 3*
- Length of one sequence *n = 10*

Step 1: Random projections

- From an l-mer to a k-mer

  Choice of k? -> **k < 3** -> **k = 2**
- Randomly chose the *k* positions in an l-mer

  Via a bitmap of 1-s and 0-s

  In the bitmap we need *k* 1-s and *l-k* 0-s

  For example: „00101“

  Use the bitmap on the sequences with a GenericShape
- Hash the k-mers
- Save them into buckets

Example

Planted motif: AGCTC

Given sequences:

CCGCGAGCTC

AGATCGTAAC

TGGGCTACCT

Example

Step 1: Random projections

Planted motif: AGCTC

Given sequences:

CCGCGAGCTC

AGATCGTAAC

TGGGCTACCT

CCGCG

00101

Use the bitmap on the 0th

l-mer of the first sequence

Example

Step 1: Random projections

Planted motif: AGCTC

Given sequences:

CCGCGAGCTC

AGATCGTAAC

TGGGCTACCT

CC**G**C**G** *l-mer*

00**101** *bitmap*

Example

Step 1: Random projections

Planted motif: AGCTC

Given sequences:

CCGCGAGCTC

AGATCGTAAC

TGGGCTACCT

CC**G**C**G** *l-mer*

00**101** *bitmap*

- - G - G *k-mer*

Example

Step 1: Random projections

Planted motif: AGCTC

Given sequences:

CCGCGAGCTC

AGATCGTAAC

TGGGCTACCT

CC**G**C**G** *l-mer*

00**101** *bitmap*

- - G - G *k-mer*

hash(**- - G - G**) = 6 *hash value*

Example

Step 1: Random projections

Planted motif: AGCTC

Given sequences:

CCGCGAGCTC

AGATCGTAAC

TGGGCTACCT

CC**G**C**G**	*l-mer*
00**1**0**1**	*bitmap*
- - G - G	*k-mer*
hash(**- - G - G**) = 6	*hash value*

Save into a bucket:

Map with [key : value]

Example

Step 1: Random projections

Planted motif: AGCTC

Given sequences:

CCGCGAGCTC

AGATCGTAAC

TGGGCTACCT

CC**G**C**G** *l-mer*

00**1**0**1** *bitmap*

- - G - G *k-mer*

hash(**- - G - G**) = 6 *hash value*

Save into a bucket:

Map with [key : value]

buckets

Example

Step 1: Random projections

Planted motif: AGCTC

Given sequences:

CCGCGAGCTC

AGATCGTAAC

TGGGCTACCT

CC**G**C**G** *l-mer*

00**101** *bitmap*

- - G - G *k-mer*

hash(**- - G - G**) = 6 *hash value*

Save into a bucket:

Map with [key : value] buckets

*The **key** is the hash value* 6

*The **value** is a vector of pairs* {[0, 0]}

Example

Step 1: Random projections

Planted motif: AGCTC

Given sequences:

CCGCGAGCTC

AGATCGTAAC

TGGGCTACCT

CC**G**C**G** *l-mer*

00**1**0**1** *bitmap*

- - G - G *k-mer*

hash(**- - G - G**) = 6 *hash value*

Save into a bucket:

Map with [key : value] buckets

The key is the hash value 6

The value is a vector of pairs {[0, 0]} *only 1 pair for now*

Example

Step 1: Random projections

Planted motif: AGCTC

Given sequences:

0 **CCGCG**AGCTC

1 AGATCGTAAC

2 TGGGCTACCT

CC**G**C**G** *l-mer*

00**1**0**1** *bitmap*

- - G - G *k-mer*

hash(**- - G - G**) = 6 *hash value*

Save into a bucket:

Map with [key : value] buckets

The key is the hash value 6

The value is a vector of pairs {[0, 0]} *only 1 pair for now*

the number of the sequence

Example

Step 1: Random projections

Planted motif: AGCTC

Given sequences:

0 1 2 3 4 5 6 7 8 9

0 **CCGCG**AGCTC

1 AGATCGTAAC

2 TGGGCTACCT

CC**G**C**G** *l-mer*

00**1**0**1** *bitmap*

- - G - G *k-mer*

hash(**- - G - G**) = 6 *hash value*

Save into a bucket:

Map with [key : value] buckets

The key is the hash value 6

The value is a vector of pairs {[0, 0]} *only 1 pair for now*

the number of the sequence

the starting position of the l-mer

Example

Step 1: Random projections

Planted motif: AGCTC

Given sequences:

C**CGCG**AGCTC

AGATCGTAAC

TGGGCTACCT

CG**C**G**A** *l-mer*

00**101** *bitmap*

Use the bitmap on the 1st

l-mer of the first sequence

Example

Step 1: Random projections

Planted motif: AGCTC

Given sequences:

C**CGCGA**GCTC

AGATCGTAAC

TGGGCTACCT

CG**C**G**A** *l-mer*

00**101** *bitmap*

- - C - A *k-mer*

Example

Step 1: Random projections

Planted motif: AGCTC

Given sequences:

C**CGCGA**GCTC

AGATCGTAAC

TGGGCTACCT

CG**C**G**A** *l-mer*

00**101** *bitmap*

- - C - A *k-mer*

hash(**- - C - A**) = 5 *hash value*

Example

Step 1: Random projections

Planted motif: AGCTC

Given sequences:

```
   0 1 2 3 4 5 6 7 8 9
0  CCGCGAGCTC
1  AGATCGTAAC
2  TGGGCTACCT
```

CG**C**G**A**	*l-mer*
00**101**	*bitmap*
- - C - A	*k-mer*
hash(**- - C - A**) = 5	*hash value*

Save into a bucket:

Map with [key : value]	buckets	
The key is the hash value	5	
The value is a vector of pairs	{[0, 1]}	*only 1 pair for now*

the number of the sequence

the starting position of the l-mer

Example

Step 1: Random projections

Buckets:

Key	*Value*	*Key*	*Value*
0	{[1, 0]}	8	{[1, 5]}
1	{[0, 5], [1, 2]}	9	{[2, 2]}
4	{[0, 3], [2, 4]}	10	{[0, 2], [0, 4], [2, 1]}
5	{[0, 1]}	11	{[1, 1], [2, 3]}
6	{[0, 0]}	13	{[2, 5]}
7	{[1, 4]}	14	{[1, 3], [2, 0]}

Example

Step 2: Refinement

Buckets:

Key	*Value*
0	: {[1, 0]}
1	: {[0, 5], [1, 2]}
4	: {[0, 3], [2, 4]}
5	: {[0, 1]}
6	: {[0, 0]}
7	: {[1, 4]}

Key	*Value*
8	: {[1, 5]}
9	: {[2, 2]}
10	: {[0, 2], [0, 4], [2, 1]}
11	: {[1, 1], [2, 3]}
13	: {[2, 5]}
14	: {[1, 3], [2, 0]}

Explore each bucket with at least *s* elements

Example

Step 2: Refinement

Buckets:

Key	*Value*
0	: {[1, 0]}
1	: {[0, 5], [1, 2]}
4	: {[0, 3], [2, 4]}
5	: {[0, 1]}
6	: {[0, 0]}
7	: {[1, 4]}
8	: {[1, 5]}
9	: {[2, 2]}
10	: {[0, 2], [0, 4], [2, 1]}
11	: {[1, 1], [2, 3]}
13	: {[2, 5]}
14	: {[1, 3], [2, 0]}

Explore each bucket with at least *s* elements

s = 2

Example

Step 2: Refinement

Buckets:

Key	*Value*
0	: {[1, 0]}
1	: {[0, 5], [1, 2]}
4	: {[0, 3], [2, 4]}
5	: {[0, 1]}
6	: {[0, 0]}
7	: {[1, 4]}
8	: {[1, 5]}
9	: {[2, 2]}
10	: {[0, 2], [0, 4], [2, 1]}
11	: {[1, 1], [2, 3]}
13	: {[2, 5]}
14	: {[1, 3], [2, 0]}

Explore each bucket with at least *s* elements

s = 2

Example

Step 2: Refinement

Buckets:

Key	*Value*
1	: {[0, 5], [1, 2]}
4	: {[0, 3], [2, 4]}
10	: {[0, 2], [0, 4], [2, 1]}
11	: {[1, 1], [2, 3]}
14	: {[1, 3], [2, 0]}

EM - Algorithm

For each bucket h

- Create an initial weight matrix Wh
- Create a position matrix given the weight matrix
- Refine the weight matrix given the position matrix
- Refine the position matrix given the new weight matrix

…

…

- Until convergence

Example

Step 2: Refinement

Create an initial weight matrix Wh for each bucket h

```
  0 1 2 3 4 5 6 7 8 9
0 C C G C G A G C T C
1 A G A T C G T A A C
2 T G G G C T A C C T
```

Example

Step 2: Refinement

Create an initial weight matrix Wh for each bucket h

For bucket 1 : {[0, 5], [1, 2]}

```
    0 1 2 3 4 5 6 7 8 9
0   C C G C G A G C T C
1   A G A T C G T A A C
2   T G G G C T A C C T
```

Example

Step 2: Refinement

Create an initial weight matrix Wh for each bucket h

For bucket 1 : {[0, 5], [1, 2]}

```
   0 1 2 3 4 5 6 7 8 9
0  CCGCGAGCTC
1  AGATCGTAAC
2  TGGGCTACCT
```

Set Wh(i, j) to be the frequency of base i among the jth positions of all l-mers in h.

Example

Step 2: Refinement

Create an initial weight matrix Wh for each bucket h

For bucket 1 : {**[0, 5]**, [1, 2]}

0 1 2 3 4 **5** 6 7 8 9

0 CCGCG**AGCTC**

1 AGATCGTAAC

2 TGGGCTACCT

Set Wh(i, j) to be the frequency of base i among the jth positions of all l-mers in h.

Example

Step 2: Refinement

Create an initial weight matrix Wh for each bucket h

For bucket 1 : {**[0, 5]**, [1, 2]}

0 1 2 3 4 **5** 6 7 8 9

0 CCGCG**AGCTC** → AGCTC

1 AGATCGTAAC

2 TGGGCTACCT

Set Wh(i, j) to be the frequency of base i among the jth positions of all l-mers in h.

Example

Step 2: Refinement

Create an initial weight matrix Wh for each bucket h

For bucket 1 : {**[0, 5]**, **[1, 2]**}

0 1 2 3 4 **5** 6 7 8 9

0 CCGCG**AGCTC** → AGCTC

0 1 **2** 3 4 5 6 7 8 9

1 AG**ATCGT**AAC

2 TGGGCTACCT

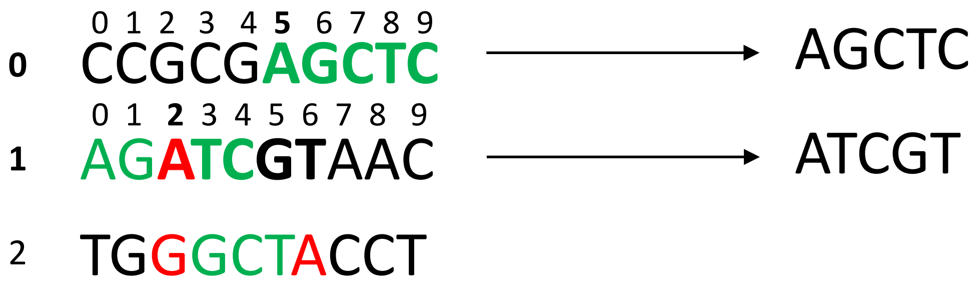
Set Wh(i, j) to be the frequency of base i among the jth positions of all l-mers in h.

Example

Step 2: Refinement

Create an initial weight matrix Wh for each bucket h

For bucket 1 : {**[0, 5]**, **[1, 2]**}



0 1 2 3 4 **5** 6 7 8 9

0 CCGCG**AGCTC** → AGCTC

0 1 **2** 3 4 5 6 7 8 9

1 AG**ATCGT**AAC → ATCGT

2 TGGGCTACCT

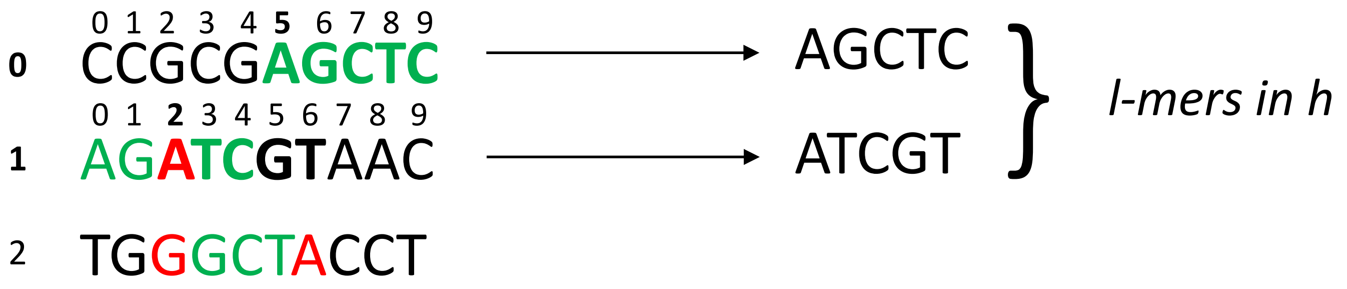
Set Wh(i, j) to be the frequency of base i among the jth positions of all l-mers in h.

Example

Step 2: Refinement

Create an initial weight matrix Wh for each bucket h

For bucket 1 : {**[0, 5]**, **[1, 2]**}



0 1 2 3 4 **5** 6 7 8 9

0 CCGCG**AGCTC**

0 1 **2** 3 4 5 6 7 8 9

1 AG**ATCGT**AAC

2 TGGGCTACCT

AGCTC

ATCGT

l-mers in h

Set Wh(i, j) to be the frequency of base i among the jth positions of all l-mers in h.

Example

Step 2: Refinement

Create an initial weight matrix Wh for each bucket h

For bucket 1 : {**[0, 5]**, **[1, 2]**}

l-mers

	0	**1**	**2**	**3**	**4**
0	A	G	C	T	C
1	A	T	C	G	T

Set Wh(i, j) to be the frequency of base i among the jth positions of all l-mers in h.

Wh

		0	**1**	**2**	**3**	**4**
A	**0**					
C	**1**					
G	**2**					
T	**3**					

Example

Step 2: Refinement

Create an initial weight matrix Wh for each bucket h

For bucket 1 : {**[0, 5]**, **[1, 2]**}

l-mers

	0	**1**	**2**	**3**	**4**
0	A	G	C	T	C
1	A	T	C	G	T

j

Wh

i

		0	**1**	**2**	**3**	**4**
A	**0**	0	0	0	0	0
C	**1**	0	0	0	0	0
G	**2**	0	0	0	0	0
T	**3**	0	0	0	0	0

Set Wh(i, j) to be the frequency of base i among the jth positions of all l-mers in h.

Example

Step 2: Refinement

Create an initial weight matrix Wh for each bucket h

For bucket 1 : {**[0, 5]**, **[1, 2]**}

l-mers

	0	**1**	**2**	**3**	**4**
0	A	G	C	T	C
1	A	T	C	G	T

Set Wh(i, j) to be the frequency of base i among the jth positions of all l-mers in h.

j →

i ↓

Wh

		0	**1**	**2**	**3**	**4**
A	**0**	0 + 1	0	0	0	0
C	**1**	0	0	0	0	0
G	**2**	0	0	0	0	0
T	**3**	0	0	0	0	0

Example

Step 2: Refinement

Create an initial weight matrix Wh for each bucket h

For bucket 1 : {**[0, 5]**, **[1, 2]**}

l-mers

	0	**1**	**2**	**3**	**4**
0	A	G	C	T	C
1	A	T	C	G	T

j →

Wh

i ↓

		0	**1**	**2**	**3**	**4**
A	**0**	1	0	0	0	0
C	**1**	0	0	0	0	0
G	**2**	0	0 + 1	0	0	0
T	**3**	0	0	0	0	0

Set Wh(i, j) to be the frequency of base i among the jth positions of all l-mers in h.

Example

Step 2: Refinement

Create an initial weight matrix Wh for each bucket h

For bucket 1 : {**[0, 5]**, **[1, 2]**}

l-mers

	0	**1**	**2**	**3**	**4**
0	A	G	C	T	C
1	A	T	C	G	T

j

Wh

i

		0	**1**	**2**	**3**	**4**
A	**0**	1	0	0	0	0
C	**1**	0	0	0 + 1	0	0
G	**2**	0	1	0	0	0
T	**3**	0	0	0	0	0

Set Wh(i, j) to be the frequency of base i among the jth positions of all l-mers in h.

Example

Step 2: Refinement

Create an initial weight matrix Wh for each bucket h

For bucket 1 : {**[0, 5]**, **[1, 2]**}

l-mers

	0	**1**	**2**	**3**	**4**
0	A	G	C	T	C
1	A	T	C	G	T

j →

Wh

i ↓

		0	**1**	**2**	**3**	**4**
A	**0**	1	0	0	0	0
C	**1**	0	0	1	0	0
G	**2**	0	1	0	0	0
T	**3**	0	0	0	0 + 1	0

Set Wh(i, j) to be the frequency of base i among the jth positions of all l-mers in h.

Example

Step 2: Refinement

Create an initial weight matrix Wh for each bucket h

For bucket 1 : {**[0, 5]**, **[1, 2]**}

l-mers

	0	**1**	**2**	**3**	**4**
0	A	G	C	T	C
1	A	T	C	G	T

j

i

Wh

		0	**1**	**2**	**3**	**4**
A	**0**	1	0	0	0	0
C	**1**	0	0	1	0	0 + 1
G	**2**	0	1	0	0	0
T	**3**	0	0	0	1	0

Set Wh(i, j) to be the frequency of base i among the jth positions of all l-mers in h.

Example

Step 2: Refinement

Create an initial weight matrix Wh for each bucket h

For bucket 1 : {**[0, 5]**, **[1, 2]**}

l-mers

	0	**1**	**2**	**3**	**4**
0	A	G	C	T	C
1	A	T	C	G	T

j →

Wh

i ↓

		0	**1**	**2**	**3**	**4**
A	**0**	1 + 1	0	0	0	0
C	**1**	0	0	1	0	1
G	**2**	0	1	0	0	0
T	**3**	0	0	0	1	0

Set Wh(i, j) to be the frequency of base i among the jth positions of all l-mers in h.

Example

Step 2: Refinement

Create an initial weight matrix Wh for each bucket h

For bucket 1 : {**[0, 5]**, **[1, 2]**}

l-mers

	0	**1**	**2**	**3**	**4**
0	A	G	C	T	C
1	A	T	C	G	T

j

Wh

i

		0	**1**	**2**	**3**	**4**
A	**0**	2	0	0	0	0
C	**1**	0	0	1	0	1
G	**2**	0	1	0	0	0
T	**3**	0	0 + 1	0	1	0

Set Wh(i, j) to be the frequency of base i among the jth positions of all l-mers in h.

Example

Step 2: Refinement

Create an initial weight matrix Wh for each bucket h

For bucket 1 : {**[0, 5]**, **[1, 2]**}

l-mers

	0	**1**	**2**	**3**	**4**
0	A	G	C	T	C
1	A	T	C	G	T

j

Wh

i

		0	**1**	**2**	**3**	**4**
A	**0**	2	0	0	0	0
C	**1**	0	0	1 + 1	0	1
G	**2**	0	1	0	0	0
T	**3**	0	1	0	1	0

Set Wh(i, j) to be the frequency of base i among the jth positions of all l-mers in h.

Example

Step 2: Refinement

Create an initial weight matrix Wh for each bucket h

For bucket 1 : {**[0, 5]**, **[1, 2]**}

l-mers

	0	**1**	**2**	**3**	**4**
0	A	G	C	T	C
1	A	T	C	G	T

j →

i ↓

Wh

		0	**1**	**2**	**3**	**4**
A	**0**	2	0	0	0	0
C	**1**	0	0	2	0	1
G	**2**	0	1	0	0 + 1	0
T	**3**	0	1	0	1	0

Set Wh(i, j) to be the frequency of base i among the jth positions of all l-mers in h.

Example

Step 2: Refinement

Create an initial weight matrix Wh for each bucket h

For bucket 1 : {**[0, 5]**, **[1, 2]**}

l-mers

	0	**1**	**2**	**3**	**4**
0	A	G	C	T	C
1	A	T	C	G	T

j →

Wh

i ↓

		0	**1**	**2**	**3**	**4**
A	**0**	2	0	0	0	0
C	**1**	0	0	2	0	1
G	**2**	0	1	0	1	0
T	**3**	0	1	0	1	0 + 1

Set Wh(i, j) to be the frequency of base i among the jth positions of all l-mers in h.

Example

Step 2: Refinement

Create an initial weight matrix Wh for each bucket h

For bucket 1 : {**[0, 5]**, **[1, 2]**}

l-mers

	0	**1**	**2**	**3**	**4**
0	A	G	C	T	C
1	A	T	C	G	T

j →

Wh

i ↓

		0	**1**	**2**	**3**	**4**
A	**0**	2	0	0	0	0
C	**1**	0	0	2	0	1
G	**2**	0	1	0	1	0
T	**3**	0	1	0	1	1

Set Wh(i, j) to be the frequency of base i among the jth positions of all l-mers in h.

To get the relative frequency divide by the number of l-mers in the bucket.

Example

Step 2: Refinement

Create an initial weight matrix Wh for each bucket h

For bucket 1 : {**[0, 5]**, **[1, 2]**}

l-mers

	0	**1**	**2**	**3**	**4**
0	A	G	C	T	C
1	A	T	C	G	T

Set Wh(i, j) to be the frequency of base i among the jth positions of all l-mers in h.

j →

i ↓

Wh

		0	**1**	**2**	**3**	**4**
A	**0**	2/2	0/2	0/2	0/2	0/2
C	**1**	0/2	0/2	2/2	0/2	1/2
G	**2**	0/2	1/2	0/2	1/2	0/2
T	**3**	0/2	1/2	0/2	1/2	1/2

To get the relative frequency divide by the number of l-mers in the bucket.

Example

Step 2: Refinement

Create an initial weight matrix Wh for each bucket h

l-mers

	0	1	2	3	4
0	A	G	C	T	C
1	A	T	C	G	T

For bucket 1 : {**[0, 5]**, **[1, 2]**}

Set Wh(i, j) to be the frequency of base i among the jth positions of all l-mers in h.

j →

Wh

i ↓

		0	**1**	**2**	**3**	**4**
A	**0**	1	0	0	0	0
C	**1**	0	0	1	0	0.5
G	**2**	0	0.5	0	0.5	0
T	**3**	0	0.5	0	0.5	0.5

To get the relative frequency divide by the number of l-mers in the bucket.

Example

Step 2: Refinement

Create an initial weight matrix Wh for each bucket h

For bucket 1 : {**[0, 5]**, **[1, 2]**}

l-mers

	0	**1**	**2**	**3**	**4**
0	A	G	C	T	C
1	A	T	C	G	T

j

Wh

i

		0	**1**	**2**	**3**	**4**
A	**0**	1 + 0.25	0 + 0.25	0 + 0.25	0 + 0.25	0 + 0.25
C	**1**	0 + 0.25	0 + 0.25	1 + 0.25	0 + 0.25	0.5 + 0.25
G	**2**	0 + 0.25	0.5 + 0.25	0 + 0.25	0.5 + 0.25	0 + 0.25
T	**3**	0 + 0.25	0.5 + 0.25	0 + 0.25	0.5 + 0.25	0.5 + 0.25

Set Wh(i, j) to be the frequency of base i among the jth positions of all l-mers in h.

Laplace correction: to avoid having probability 0 add a background probability.

Example

Step 2: Refinement

Create an initial weight matrix Wh for each bucket h

For bucket 1 : {**[0, 5]**, **[1, 2]**}

l-mers

	0	**1**	**2**	**3**	**4**
0	A	G	C	T	C
1	A	T	C	G	T

Set Wh(i, j) to be the frequency of base i among the jth positions of all l-mers in h.

j

Wh

i

		0	**1**	**2**	**3**	**4**
A	**0**	1.25	0.25	0.25	0.25	0.25
C	**1**	0.25	0.25	1.25	0.25	0.75
G	**2**	0.25	0.75	0.25	0.75	0.25
T	**3**	0.25	0.75	0.25	0.75	0.75

Laplace correction: to avoid having probability 0 add a background probability.

Example

Step 2: Refinement

Create a position matrix given the weight matrix

sequences

	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

.
.

Wh

		0	**1**	**2**	**3**	**4**
A	**0**	1.25	0.25	0.25	0.25	0.25
C	**1**	0.25	0.25	1.25	0.25	0.75
G	**2**	0.25	0.75	0.25	0.75	0.25
T	**3**	0.25	0.75	0.25	0.75	0.75

posM

	0	**1**	**2**	**3**	**4**	**5**
0						
1						
2						

$$\mathrm{posM}'(\mathrm{i},\mathrm{j}) = \Pr\left(z_{ij} = 1 \,\middle|\, \mathrm{sequences}, \mathrm{Wh}\right)$$

$$= \frac{\Pr\left(\mathrm{sequences} \,\middle|\, z_{ij} = 1, \mathrm{Wh}\right)}{\sum_{k=0}^{4} \Pr(\mathrm{sequences}\; |z_{ik} = 1, \mathrm{Wh})}$$

where $z_{ij} = 1$ means that j is the starting position of the motif in sequence i

Example

Step 2: Refinement

Create a position matrix given the weight matrix

sequences

	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

⋮

posM

	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0

Wh

		0	1	2	3	4
A	**0**	1.25	0.25	0.25	0.25	0.25
C	**1**	0.25	0.25	1.25	0.25	0.75
G	**2**	0.25	0.75	0.25	0.75	0.25
T	**3**	0.25	0.75	0.25	0.75	0.75

$$\mathrm{posM}'(\mathrm{i},\mathrm{j}) = \Pr\left(z_{ij} = 1 \,\middle|\, \text{sequences}, \mathrm{Wh}\right)$$

$$= \frac{\Pr\left(\text{sequences} \,\middle|\, z_{ij} = 1, \mathrm{Wh}\right)}{\sum_{k=0}^{4} \Pr(\text{sequences} \mid z_{ik} = 1, \mathrm{Wh})}$$

where $z_{ij} = 1$ means that j is the starting position of the motif in sequence i

Example

Step 2: Refinement

Create a position matrix given the weight matrix

sequences

	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

⋮

Wh

		0	1	2	3	4
A	**0**	1.25	0.25	0.25	0.25	0.25
C	**1**	0.25	0.25	1.25	0.25	0.75
G	**2**	0.25	0.75	0.25	0.75	0.25
T	**3**	0.25	0.75	0.25	0.75	0.75

posM

	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0

$$\mathrm{posM}'(0,0) = \Pr(z_{00} = 1 \mid \text{sequences}, \mathrm{Wh})$$

$$= \frac{\Pr(\text{sequences} \mid z_{00} = 1, \mathrm{Wh})}{\sum_{k=0}^{4} \Pr(\text{sequences} \mid z_{0k} = 1, \mathrm{Wh})}$$

where $z_{00} = 1$ means that 0 is the starting position of the motif in sequence 0

Example

Step 2: Refinement

Create a position matrix given the weight matrix

sequences

	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

posM

	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0

Wh

		0	1	2	3	4
A	**0**	1.25	0.25	0.25	0.25	0.25
C	**1**	0.25	0.25	1.25	0.25	0.75
G	**2**	0.25	0.75	0.25	0.75	0.25
T	**3**	0.25	0.75	0.25	0.75	0.75

$$\mathrm{posM}'(0,0) = \frac{\Pr(\text{sequences} \mid z_{00} = 1, \mathrm{Wh})}{\sum_{k=0}^{4} \Pr(\text{sequences } | z_{0k} = 1, \mathrm{Wh})}$$

$$=$$

Example

Step 2: Refinement

Create a position matrix given the weight matrix

sequences

	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

posM

	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0

Wh

		0	1	2	3	4
A	**0**	1.25	0.25	0.25	0.25	0.25
C	**1**	**0.25**	**0.25**	1.25	**0.25**	0.75
G	**2**	0.25	0.75	**0.25**	0.75	**0.25**
T	**3**	0.25	0.75	0.25	0.75	0.75

$$\mathrm{posM}'(0,0) = \frac{\Pr(\mathrm{sequences} \mid z_{00} = 1, \mathrm{Wh})}{\sum_{k=0}^{4} \Pr(\mathrm{sequences} \mid z_{0k} = 1, \mathrm{Wh})}$$

$$= \frac{0.25*0.25*0.25*0.25*0.25}{}$$

Example

Step 2: Refinement

Create a position matrix given the weight matrix

sequences

	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

posM

	0	**1**	**2**	**3**	**4**	**5**
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0

Wh

		0	**1**	**2**	**3**	**4**
A	**0**	1.25	0.25	0.25	0.25	0.25
C	**1**	**0.25**	**0.25**	1.25	**0.25**	0.75
G	**2**	0.25	0.75	**0.25**	0.75	**0.25**
T	**3**	0.25	0.75	0.25	0.75	0.75

$$\mathrm{posM}'(0,0) = \frac{\Pr(\text{sequences} \mid z_{00} = 1, \mathrm{Wh})}{\sum_{k=0}^{4} \Pr(\text{sequences} \mid z_{0k} = 1, \mathrm{Wh})}$$

$$= \frac{0.0009765625}{\qquad\qquad}$$

Example

Step 2: Refinement

Create a position matrix given the weight matrix

sequences

	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

posM

	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0

Wh

		0	1	2	3	4
A	**0**	1.25	0.25	0.25	0.25	0.25
C	**1**	**0.25**	**0.25**	1.25	**0.25**	0.75
G	**2**	0.25	0.75	**0.25**	0.75	**0.25**
T	**3**	0.25	0.75	0.25	0.75	0.75

$$\text{posM}'(0,0) = \frac{\Pr(\text{sequences} \mid z_{00} = 1, \text{Wh})}{\sum_{k=0}^{4} \Pr(\text{sequences} \mid z_{0k} = 1, \text{Wh})}$$

$$= \frac{0.0009765625}{0.0009765625 +}$$

Example

Step 2: Refinement

Create a position matrix given the weight matrix

sequences

	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

posM

	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0

Wh

		0	1	2	3	4
A	**0**	1.25	0.25	0.25	0.25	**0.25**
C	**1**	**0.25**	0.25	**1.25**	0.25	0.75
G	**2**	0.25	**0.75**	0.25	**0.75**	0.25
T	**3**	0.25	0.75	0.25	0.75	0.75

$$\text{posM}'(0,0) = \frac{\Pr(\text{sequences} \mid z_{00} = 1, \text{Wh})}{\sum_{k=0}^{4} \Pr(\text{sequences} \mid z_{0k} = 1, \text{Wh})}$$

$$= \frac{0.0009765625}{0.0009765625 + 0.25*0.75*1.25*0.75*0.25}$$

Example

Step 2: Refinement

Create a position matrix given the weight matrix

sequences

	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

posM

	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0

Wh

		0	1	2	3	4
A	**0**	1.25	0.25	0.25	0.25	**0.25**
C	**1**	**0.25**	0.25	**1.25**	0.25	0.75
G	**2**	0.25	**0.75**	0.25	**0.75**	0.25
T	**3**	0.25	0.75	0.25	0.75	0.75

$$\mathrm{posM}'(0,0) = \frac{\Pr(\text{sequences} \mid z_{00} = 1, \mathrm{Wh})}{\sum_{k=0}^{4} \Pr(\text{sequences} \mid z_{0k} = 1, \mathrm{Wh})}$$

$$= \frac{0.0009765625}{0.0009765625 + 0.0439453125}$$

Example

Step 2: Refinement

Create a position matrix given the weight matrix

sequences

	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

posM

	0	**1**	**2**	**3**	**4**	**5**
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0

Wh

		0	**1**	**2**	**3**	**4**
A	**0**	1.25	0.25	0.25	0.25	**0.25**
C	**1**	**0.25**	0.25	**1.25**	0.25	0.75
G	**2**	0.25	**0.75**	0.25	**0.75**	0.25
T	**3**	0.25	0.75	0.25	0.75	0.75

$$\mathrm{posM}'(0,0) = \frac{\Pr(\text{sequences} \mid z_{00} = 1, \mathrm{Wh})}{\sum_{k=0}^{4} \Pr(\text{sequences} \mid z_{0k} = 1, \mathrm{Wh})}$$

$$= \frac{0.0009765625}{0.044921875 + }$$

Example

Step 2: Refinement

Create a position matrix given the weight matrix

sequences

	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

posM

	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0

Wh

		0	1	2	3	4
A	**0**	1.25	0.25	0.25	**0.25**	0.25
C	**1**	0.25	**0.25**	1.25	0.25	0.75
G	**2**	**0.25**	0.75	**0.25**	0.75	**0.25**
T	**3**	0.25	0.75	0.25	0.75	0.75

$$\mathrm{posM}'(0,0) = \frac{\Pr(\text{sequences} \mid z_{00} = 1, \mathrm{Wh})}{\sum_{k=0}^{4} \Pr(\text{sequences} \mid z_{0k} = 1, \mathrm{Wh})}$$

$$= \frac{0.0009765625}{0.044921875 + 0.0009765625}$$

Example

Step 2: Refinement

Create a position matrix given the weight matrix

sequences

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

posM

	0	**1**	**2**	**3**	**4**	**5**
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0

Wh

		0	**1**	**2**	**3**	**4**
A	**0**	1.25	0.25	**0.25**	0.25	0.25
C	**1**	**0.25**	0.25	1.25	0.25	**0.75**
G	**2**	0.25	**0.75**	0.25	**0.75**	0.25
T	**3**	0.25	0.75	0.25	0.75	0.75

$$\mathrm{posM}'(0,0) = \frac{\Pr(\mathrm{sequences} \mid z_{00} = 1, \mathrm{Wh})}{\sum_{k=0}^{4} \Pr(\mathrm{sequences} \mid z_{0k} = 1, \mathrm{Wh})}$$

$$= \frac{0.0009765625}{0.0458984375 + 0.0263671875 +}$$

Example

Step 2: Refinement

Create a position matrix given the weight matrix

sequences

	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

posM

	0	**1**	**2**	**3**	**4**	**5**
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0

Wh

		0	**1**	**2**	**3**	**4**
A	**0**	1.25	**0.25**	0.25	0.25	0.25
C	**1**	0.25	0.25	1.25	**0.25**	0.75
G	**2**	**0.25**	0.75	**0.25**	0.75	0.25
T	**3**	0.25	0.75	0.25	0.75	**0.75**

$$\mathrm{posM}'(0,0) = \frac{\Pr(\text{sequences} \mid z_{00} = 1, \mathrm{Wh})}{\sum_{k=0}^{4} \Pr(\text{sequences} \mid z_{0k} = 1, \mathrm{Wh})}$$

$$= \frac{0.0009765625}{0.0458984375 + 0.0263671875 + 0.0029296875 +}$$

Example

Step 2: Refinement

Create a position matrix given the weight matrix

sequences

	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

Wh

		0	1	2	3	4
A	**0**	**1.25**	0.25	0.25	0.25	0.25
C	**1**	0.25	0.25	**1.25**	0.25	**0.75**
G	**2**	0.25	**0.75**	0.25	0.75	0.25
T	**3**	0.25	0.75	0.25	**0.75**	0.75

posM

	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0

$$\mathrm{posM}'(0,0) = \frac{\Pr(\text{sequences} \mid z_{00} = 1, \mathrm{Wh})}{\sum_{k=0}^{4} \Pr(\text{sequences} \mid z_{0k} = 1, \mathrm{Wh})}$$

$$= \frac{0.0009765625}{0.0458984375 + 0.0263671875 + 0.0029296875 + 0.6591796875}$$

Example

Step 2: Refinement

Create a position matrix given the weight matrix

sequences

	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

Wh

		0	**1**	**2**	**3**	**4**
A	**0**	**1.25**	0.25	0.25	0.25	0.25
C	**1**	0.25	0.25	**1.25**	0.25	**0.75**
G	**2**	0.25	**0.75**	0.25	0.75	0.25
T	**3**	0.25	0.75	0.25	**0.75**	0.75

posM

	0	**1**	**2**	**3**	**4**	**5**
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0

⋮

$$\mathrm{posM}'(0,0) = \frac{\Pr(\text{sequences} \mid z_{00} = 1, \mathrm{Wh})}{\sum_{k=0}^{4} \Pr(\text{sequences} \mid z_{0k} = 1, \mathrm{Wh})}$$

$$= \frac{0.0009765625}{0.734375}$$

Example

Step 2: Refinement

Create a position matrix given the weight matrix

sequences

	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

Wh

		0	1	2	3	4
A	**0**	**1.25**	0.25	0.25	0.25	0.25
C	**1**	0.25	0.25	**1.25**	0.25	**0.75**
G	**2**	0.25	**0.75**	0.25	0.75	0.25
T	**3**	0.25	0.75	0.25	**0.75**	0.75

posM

	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0

$$\mathrm{posM}'(0,0) = \frac{\Pr(\text{sequences} \mid z_{00} = 1, \mathrm{Wh})}{\sum_{k=0}^{4} \Pr(\text{sequences} \mid z_{0k} = 1, \mathrm{Wh})}$$

$$= 0.001329787$$

Example

Step 2: Refinement

Create a position matrix given the weight matrix

sequences

	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

Wh

		0	1	2	3	4
A	**0**	1.25	0.25	0.25	0.25	0.25
C	**1**	0.25	0.25	1.25	0.25	0.75
G	**2**	0.25	0.75	0.25	0.75	0.25
T	**3**	0.25	0.75	0.25	0.75	0.75

posM

	0	1	2	3	4	5
0	0.001329	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0

⋮

$$\mathrm{posM}'(0,0) = \frac{\Pr(\text{sequences} \mid z_{00} = 1, \mathrm{Wh})}{\sum_{k=0}^{4} \Pr(\text{sequences} \mid z_{0k} = 1, \mathrm{Wh})}$$

$$= 0.001329787$$

Example

Step 2: Refinement

Create a position matrix given the weight matrix

posM

	0	**1**	**2**	**3**	**4**	**5**
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

sequences

	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

posM

	0	1	2	3	4	5
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

Wh

		0	1	2	3	4
A	0	1.25	0.25	0.25	0.25	0.25
C	1	0.25	0.25	1.25	0.25	0.75
G	2	0.25	0.75	0.25	0.75	0.25
T	3	0.25	0.75	0.25	0.75	0.75

→

W

		0	1	2	3	4
A	0					
C	1					
G	2					
T	3					

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

posM

	0	1	2	3	4	5
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

sequences

	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

Wh

		0	1	2	3	4
A	**0**	1.25	0.25	0.25	0.25	0.25
C	**1**	0.25	0.25	1.25	0.25	0.75
G	**2**	0.25	0.75	0.25	0.75	0.25
T	**3**	0.25	0.75	0.25	0.75	0.75

W

		0	1	2	3	4
A	**0**	0	0	0	0	0
C	**1**	0	0	0	0	0
G	**2**	0	0	0	0	0
T	**3**	0	0	0	0	0

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

sequences

	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

posM

	0	1	2	3	4	5
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

Wh

		0	1	2	3	4
A	**0**	1.25	0.25	0.25	0.25	0.25
C	**1**	0.25	0.25	1.25	0.25	0.75
G	**2**	0.25	0.75	0.25	0.75	0.25
T	**3**	0.25	0.75	0.25	0.75	0.75

$$W(0,0) = W_{A,0} = \frac{W'_{A,0}}{\sum_{i=A,C,G,T} W'_{i,0}}$$

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

sequences

	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

posM

	0	1	2	3	4	5
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

Wh

		0	1	2	3	4
A	**0**	1.25	0.25	0.25	0.25	0.25
C	**1**	0.25	0.25	1.25	0.25	0.75
G	**2**	0.25	0.75	0.25	0.75	0.25
T	**3**	0.25	0.75	0.25	0.75	0.75

$$W(0,0) = W_{A,0} = \frac{W'_{A,0}}{\sum_{i=A,C,G,T} W'_{i,0}}$$

The probability that A is the first letter of the motif

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

	0	1	2	3	4	5				
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

Wh

		0	1	2	3	4
A	**0**	1.25	0.25	0.25	0.25	0.25
C	**1**	0.25	0.25	1.25	0.25	0.75
G	**2**	0.25	0.75	0.25	0.75	0.25
T	**3**	0.25	0.75	0.25	0.75	0.75

posM

	0	1	2	3	4	5
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

$$W(0,0) = W_{A,0} = \frac{W'_{A,0}}{\sum_{i=A,C,G,T} W'_{i,0}}$$

← *The probability that A is the first letter of the motif*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

	0	1	2	3	4	5				
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

Wh

posM

	0	**1**	**2**	**3**	**4**	**5**
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

		0	**1**	**2**	**3**	**4**
A	**0**	1.25	0.25	0.25	0.25	0.25
C	**1**	0.25	0.25	1.25	0.25	0.75
G	**2**	0.25	0.75	0.25	0.75	0.25
T	**3**	0.25	0.75	0.25	0.75	0.75

$W'_{BASE, start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and  pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

	0	1	2	3	4	5				
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

posM

	0	1	2	3	4	5
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

Wh

		0	1	2	3	4
A	**0**	1.25	0.25	0.25	0.25	0.25
C	**1**	0.25	0.25	1.25	0.25	0.75
G	**2**	0.25	0.75	0.25	0.75	0.25
T	**3**	0.25	0.75	0.25	0.75	0.75

$W'_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and  pos=0 to seq_length - motif_length + 1
            $W[\underbrace{ordValue(sequences[i][j])}_{BASE}][\underbrace{start}_{start}] \mathrel{+}= posM[i][pos]$
```

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

j →

Refine the weight matrix given the position matrix

i ↓

	0	1	2	3	4	5				
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

posM

	0	**1**	**2**	**3**	**4**	**5**
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

Wh

		0	**1**	**2**	**3**	**4**
A	**0**	1.25	0.25	0.25	0.25	0.25
C	**1**	0.25	0.25	1.25	0.25	0.75
G	**2**	0.25	0.75	0.25	0.75	0.25
T	**3**	0.25	0.75	0.25	0.75	0.75

$W'_{BASE, start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and  pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

i

	0	1	2	3	4	5				
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

Wh

posM

j

i

	0	**1**	**2**	**3**	**4**	**5**
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

$W'_{BASE,\ start}$

		0	**1**	**2**	**3**	**4**
A	**0**	1.25	0.25	0.25	0.25	0.25
C	**1**	0.25	0.25	1.25	0.25	0.75
G	**2**	0.25	0.75	0.25	0.75	0.25
T	**3**	0.25	0.75	0.25	0.75	0.75

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and  pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

	0	1	2	3	4	5				
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

Wh

j

posM

i

	0	1	2	3	4	5
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

		0	1	2	3	4
A	**0**	1.25	0.25	0.25	0.25	0.25
C	**1**	0.25	0.25	1.25	0.25	0.75
G	**2**	0.25	0.75	0.25	0.75	0.25
T	**3**	0.25	0.75	0.25	0.75	0.75

$W'_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and  pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

	0	1	2	3	4	5				
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

posM

j

i

	0	1	2	3	4	5
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

start = 0 **j = start = 0**

i = 0 **pos = 0**

$W'_{BASE,\ start}$

W

		0	1	2	3	4
A	**0**	0	0	0	0	0
C	**1**	0	0	0	0	0
G	**2**	0	0	0	0	0
T	**3**	0	0	0	0	0

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and  pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

	0	1	2	3	4	5				
	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

posM

j

i

	0	1	2	3	4	5
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

start = 0 j = start = 0

i = 0 pos = 0

$W'_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

sequences[0][0] = „C“

W

		0	1	2	3	4
A	0	0	0	0	0	0
C	1	0	0	0	0	0
G	2	0	0	0	0	0
T	3	0	0	0	0	0

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

	0	1	2	3	4	5				
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

j

i

posM

	0	**1**	**2**	**3**	**4**	**5**
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

W

		0	**1**	**2**	**3**	**4**
A	**0**	0	0	0	0	0
C	**1**	0	0	0	0	0
G	**2**	0	0	0	0	0
T	**3**	0	0	0	0	0

start = 0 **j = start = 0**

i = 0 **pos = 0**

$W'_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

ordValue(„C“) = 1

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

	0	1	2	3	4	5				
	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

j

i

posM

	0	1	2	3	4	5
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

start = 0 j = start = 0

i = 0 pos = 0

$W'_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

W[1][0]

W

		0	1	2	3	4
A	0	0	0	0	0	0
C	1	0	0	0	0	0
G	2	0	0	0	0	0
T	3	0	0	0	0	0

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

	0	1	2	3	4	5				
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

posM

j

i

	0	**1**	**2**	**3**	**4**	**5**
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

start = 0 **j = start = 0**

i = 0 **pos = 0**

W

		0	**1**	**2**	**3**	**4**
A	**0**	0	0	0	0	0
C	**1**	0	0	0	0	0
G	**2**	0	0	0	0	0
T	**3**	0	0	0	0	0

$W'_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and  pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

W[1][0]

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j →

pos →

i ↓

	0	1	2	3	4	5				
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

j →

i ↓

posM

	0	1	2	3	4	5
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

W

		0	1	2	3	4
A	**0**	0	0	0	0	0
C	**1**	0	0	0	0	0
G	**2**	0	0	0	0	0
T	**3**	0	0	0	0	0

start = 0 **j = start = 0**

i = 0 **pos = 0**

W'$_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

W[1][0] += posM[0][0]

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

	0	1	2	3	4	5				
	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

posM

j

i

	0	1	2	3	4	5
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

W

		0	1	2	3	4
A	0	0	0	0	0	0
C	1	0.00132979	0	0	0	0
G	2	0	0	0	0	0
T	3	0	0	0	0	0

start = 0 **j = start = 0**

i = 0 **pos = 0**

W'$_{BASE, start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

W[1][0] += posM[0][0]

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

	0	1	2	3	4	5				
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

j

i

posM

	0	1	2	3	4	5
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

W

		0	1	2	3	4
A	**0**	0	0	0	0	0
C	**1**	0.00132979	0	0	0	0
G	**2**	0	0	0	0	0
T	**3**	0	0	0	0	0

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

start = 0 **j = 1**

i = 0 **pos = 1**

W'$_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

sequences[0][1] = „C“

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

	0	1	2	3	4	5				
	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

start = 0 j = 1

i = 0 pos = 1

$W'_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

ordValue(„C“) = 1

j

i

posM

	0	1	2	3	4	5
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

W

		0	1	2	3	4
A	0	0	0	0	0	0
C	1	0.00132979	0	0	0	0
G	2	0	0	0	0	0
T	3	0	0	0	0	0

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

	0	1	2	3	4	5				
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

j

i

posM

	0	**1**	**2**	**3**	**4**	**5**
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

start = 0 **j = 1**

i = 0 **pos = 1**

W

		0	**1**	**2**	**3**	**4**
A	**0**	0	0	0	0	0
C	**1**	0.00132979	0	0	0	0
G	**2**	0	0	0	0	0
T	**3**	0	0	0	0	0

$W'_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

W[1][0]

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

	0	1	2	3	4	5				
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

posM

j

i

	0	**1**	**2**	**3**	**4**	**5**
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

start = 0 **j = 1**

i = 0 **pos = 1**

$W'_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

W[1][0] += posM[0][1]

W

		0	**1**	**2**	**3**	**4**
A	**0**	0	0	0	0	0
C	**1**	0.00132979 + 0.0598404	0	0	0	0
G	**2**	0	0	0	0	0
T	**3**	0	0	0	0	0

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

	0	1	2	3	4	5				
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

start = 0 **j = 2**

i = 0 **pos = 2**

W'

$W'_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

sequences[0][1] = „G“

j

i

posM

	0	**1**	**2**	**3**	**4**	**5**
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

W

		0	**1**	**2**	**3**	**4**
A	**0**	0	0	0	0	0
C	**1**	0.06117019	0	0	0	0
G	**2**	0	0	0	0	0
T	**3**	0	0	0	0	0

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

	0	1	2	3	4	5				
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

start = 0 **j = 2**

i = 0 **pos = 2**

W‘$_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and  pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

W[2][0]

posM

j

i

	0	**1**	**2**	**3**	**4**	**5**
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

W

		0	**1**	**2**	**3**	**4**
A	**0**	0	0	0	0	0
C	**1**	0.06117019	0	0	0	0
G	**2**	0	0	0	0	0
T	**3**	0	0	0	0	0

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j →

pos →

i ↓

pos	0	1	2	3	4	5				
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

start = 0 **j = 2**

i = 0 **pos = 2**

W'$_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and  pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

W[2][0] += posM[0][2]

posM

j →

i ↓

	0	1	2	3	4	5
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

W

		0	1	2	3	4
A	**0**	0	0	0	0	0
C	**1**	0.06117019	0	0	0	0
G	**2**	0.00132979	0	0	0	0
T	**3**	0	0	0	0	0

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

pos		0	1	2	3	4	5				
j		0	1	2	3	4	5	6	7	8	9
i	0	C	C	G	C	G	A	G	C	T	C
	1	A	G	A	T	C	G	T	A	A	C
	2	T	G	G	G	C	T	A	C	C	T

start = 0 j = 3

i = 0 pos = 3

W‘$_{BASE, start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

W[1][0] += posM[0][3]

posM

i \ j	0	1	2	3	4	5
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

W

		0	1	2	3	4
A	0	0	0	0	0	0
C	1	0.09707449	0	0	0	0
G	2	0.00132979	0	0	0	0
T	3	0	0	0	0	0

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

pos	0	1	2	3	4	5				
i \ j	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

start = 0   j = 4

i = 0   pos = 4

$W'_{BASE, start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

W[2][0] += posM[0][4]

posM

i \ j	0	1	2	3	4	5
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

W

		0	1	2	3	4
A	0	0	0	0	0	0
C	1	0.09707449	0	0	0	0
G	2	0.00531915	0	0	0	0
T	3	0	0	0	0	0

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

	0	1	2	3	4	5				
	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

start = 0 j = 5

i = 0 pos = 5

$W'_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

W[0][0] += posM[0][5]

j

i

posM

	0	1	2	3	4	5
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

W

		0	1	2	3	4
A	0	0.897606	0	0	0	0
C	1	0.09707449	0	0	0	0
G	2	0.00531915	0	0	0	0
T	3	0	0	0	0	0

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

	pos 0	pos 1	pos 2	pos 3	pos 4	pos 5				
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

j

i

posM

	0	**1**	**2**	**3**	**4**	**5**
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

start = 0 **j = 0**

i = 1 **pos = 0**

W

		0	**1**	**2**	**3**	**4**
A	**0**	1.061044	0	0	0	0
C	**1**	0.09707449	0	0	0	0
G	**2**	0.00531915	0	0	0	0
T	**3**	0	0	0	0	0

W'$_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

W[0][0] += posM[1][0]

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

	0	1	2	3	4	5				
	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

posM

j

i

	0	1	2	3	4	5
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

W

		0	1	2	3	4
A	0	1.061044	0	0	0	0
C	1	0.09707449	0	0	0	0
G	2	0.0065298	0	0	0	0
T	3	0	0	0	0	0

start = 0 **j = 1**

i = 1 **pos = 1**

$W'_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and  pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

W[2][0] += posM[1][1]

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

pos	0	1	2	3	4	5				
	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

j

i

posM

	0	1	2	3	4	5
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

W

		0	1	2	3	4
A	0	1.061044	0	0	0	0
C	1	0.09707449	0	0	0	0
G	2	0.0065298	0	0	0	0
T	3	0	0	0	0	0

$W'_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and  pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

	0	1	2	3	4	5				
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

posM

j

i

	0	**1**	**2**	**3**	**4**	**5**
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

W

		0	**1**	**2**	**3**	**4**
A	**0**	1.061044	0	0	0	0
C	**1**	0.09707449	0	0	0	0
G	**2**	0.0065298	0	0	0	0
T	**3**	0	0	0	0	0

$W'_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and  pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j →

pos →

i ↓

	pos									
	0	**1**	**2**	**3**	**4**	**5**				
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

posM

j →

i ↓

	0	**1**	**2**	**3**	**4**	**5**
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

W

		0	**1**	**2**	**3**	**4**
A	**0**	1.061044	0	0	0	0
C	**1**	0.09707449	0	0	0	0
G	**2**	0.0065298	0	0	0	0
T	**3**	0	0	0	0	0

$W'_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and  pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

	0	**1**	**2**	**3**	**4**	**5**				
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

W'$_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and  pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

j

posM

i

	0	**1**	**2**	**3**	**4**	**5**
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

W

		0	**1**	**2**	**3**	**4**
A	**0**	1.061044	0	0	0	0
C	**1**	0.09707449	0	0	0	0
G	**2**	0.0065298	0	0	0	0
T	**3**	0	0	0	0	0

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

pos →	0	1	2	3	4	5				
	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

j

i

posM

	0	1	2	3	4	5
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

W

		0	1	2	3	4
A	0	1.061044	0	0	0	0
C	1	0.09707449	0	0	0	0
G	2	0.0065298	0	0	0	0
T	3	0	0	0	0	0

$W'_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and  pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

	0	1	2	3	4	5				
	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

j

i

posM

	0	1	2	3	4	5
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

W

		0	1	2	3	4
A	0	1.061044	0	0	0	0
C	1	0.09707449	0	0	0	0
G	2	0.0065298	0	0	0	0
T	3	0	0	0	0	0

$W'_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and  pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

pos	0	1	2	3	4	5				
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

j

i

posM

	0	**1**	**2**	**3**	**4**	**5**
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

W

		0	**1**	**2**	**3**	**4**
A	**0**	1.061044	0	0	0	0
C	**1**	0.09707449	0	0	0	0
G	**2**	0.0065298	0	0	0	0
T	**3**	0	0	0	0	0

$W'_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and  pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

pos	0	1	2	3	4	5				
	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

j

i

posM

	0	1	2	3	4	5
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

W

		0	1	2	3	4
A	0	1.061044	0	0	0	0
C	1	0.09707449	0	0	0	0
G	2	0.0065298	0	0	0	0
T	3	0	0	0	0	0

W'$_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and  pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

	0	1	2	3	4	5				
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

j

posM

i

	0	1	2	3	4	5
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

W

		0	1	2	3	4
A	**0**	1.061044	0	0	0	0
C	**1**	0.09707449	0	0	0	0
G	**2**	0.0065298	0	0	0	0
T	**3**	0	0	0	0	0

$W'_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

	0	1	2	3	4	5				
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

j

posM

i

	0	1	2	3	4	5
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

W

		0	1	2	3	4
A	**0**	1.061044	0	0	0	0
C	**1**	0.09707449	0	0	0	0
G	**2**	0.0065298	0	0	0	0
T	**3**	0	0	0	0	0

$W'_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and  pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

		pos 0	pos 1	pos 2	pos 3	pos 4	pos 5			
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

j

i

posM

	0	**1**	**2**	**3**	**4**	**5**
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

W

		0	**1**	**2**	**3**	**4**
A	**0**	1.061044	0	0	0	0
C	**1**	0.09707449	0	0	0	0
G	**2**	0.0065298	0	0	0	0
T	**3**	0	0	0	0	0

$W'_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and  pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

		0	1	2	3	4	5			
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

j

i

posM

	0	**1**	**2**	**3**	**4**	**5**
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

W

		0	**1**	**2**	**3**	**4**
A	**0**	1.061044	0	0	0	0
C	**1**	0.09707449	0	0	0	0
G	**2**	0.0065298	0	0	0	0
T	**3**	0	0	0	0	0

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

$W'_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and  pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

		0	1	2	3	4	5			
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

posM

j

i

	0	**1**	**2**	**3**	**4**	**5**
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

W

		0	**1**	**2**	**3**	**4**
A	**0**	1.061044	0	0	0	0
C	**1**	0.09707449	0	0	0	0
G	**2**	0.0065298	0	0	0	0
T	**3**	0	0	0	0	0

W‘$_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

		0	**1**	**2**	**3**	**4**	**5**			
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

j

i

posM

	0	**1**	**2**	**3**	**4**	**5**
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

W

		0	**1**	**2**	**3**	**4**
A	**0**	1.061044	0	0	0	0
C	**1**	0.09707449	0	0	0	0
G	**2**	0.0065298	0	0	0	0
T	**3**	0	0	0	0	0

$W'_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

		0	1	2	3	4	5			
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

j

i

posM

	0	**1**	**2**	**3**	**4**	**5**
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

W

		0	**1**	**2**	**3**	**4**
A	**0**	1.061044	0	0	0	0
C	**1**	0.09707449	0	0	0	0
G	**2**	0.0065298	0	0	0	0
T	**3**	0	0	0	0	0

$W'_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and  pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

		0	1	2	3	4	5			
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

posM

j

i

	0	**1**	**2**	**3**	**4**	**5**
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

W

		0	**1**	**2**	**3**	**4**
A	**0**	1.061044	0	0	0	0
C	**1**	0.09707449	0	0	0	0
G	**2**	0.0065298	0	0	0	0
T	**3**	0	0	0	0	0

$W'_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

		0	**1**	**2**	**3**	**4**	**5**		
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

j

i

posM

	0	**1**	**2**	**3**	**4**	**5**
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

W

		0	**1**	**2**	**3**	**4**
A	**0**	1.061044	0	0	0	0
C	**1**	0.09707449	0	0	0	0
G	**2**	0.0065298	0	0	0	0
T	**3**	0	0	0	0	0

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

$W'_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and  pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

			0	1	2	3	4	5		
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

j

posM

i

	0	**1**	**2**	**3**	**4**	**5**
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

W

		0	**1**	**2**	**3**	**4**
A	**0**	1.061044	0	0	0	0
C	**1**	0.09707449	0	0	0	0
G	**2**	0.0065298	0	0	0	0
T	**3**	0	0	0	0	0

<u>start</u> is the position in the motif

<u>i</u> is the number of the sequence

<u>j</u> is the position in the sequence

<u>pos</u> is the position in the window

$W'_{BASE, start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and  pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j

pos

i

				0	1	2	3	4	5	
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

j

i

posM

	0	1	2	3	4	5
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

W

		0	1	2	3	4
A	**0**	1.061044	0	0	0	0
C	**1**	0.09707449	0	0	0	0
G	**2**	0.0065298	0	0	0	0
T	**3**	0	0	0	0	0

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

$W'_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and  pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

j →

pos →

i ↓

					0	1	2	3	4	5
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

posM

j →

i ↓

	0	1	2	3	4	5
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

W

		0	1	2	3	4
A	**0**	1.061044	0	0	0	0
C	**1**	0.09707449	0	0	0	0
G	**2**	0.0065298	0	0	0	0
T	**3**	0	0	0	0	0

$W'_{BASE,\ start}$

```
For start = 0 to motif length
    for i=0 to number of sequences
        for j=start and  pos=0 to seq_length - motif_length + 1
            W[ordValue(sequences[i][j])][start] += posM[i][pos]
```

start *is the position in the motif*

i *is the number of the sequence*

j *is the position in the sequence*

pos *is the position in the window*

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

W

		0	1	2	3	4
A	0	W(0,0)	W(0,1)	W(0,2)	W(0,3)	W(0,4)
C	1	W(1,0)	W(1,1)	W(1,2)	W(1,3)	W(1,4)
G	2	W(2,0)	W(2,1)	W(2,2)	W(2,3)	W(2,4)
T	3	W(3,0)	W(3,1)	W(3,2)	W(3,3)	W(3,4)

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

Wh

		0	1	2	3	4
A	**0**	1.25	0.25	0.25	0.25	0.25
C	**1**	0.25	0.25	1.25	0.25	0.75
G	**2**	0.25	0.75	0.25	0.75	0.25
T	**3**	0.25	0.75	0.25	0.75	0.75

+

W

		0	1	2	3	4
A	**0**	W(0,0)	W(0,1)	W(0,2)	W(0,3)	W(0,4)
C	**1**	W(1,0)	W(1,1)	W(1,2)	W(1,3)	W(1,4)
G	**2**	W(2,0)	W(2,1)	W(2,2)	W(2,3)	W(2,4)
T	**3**	W(3,0)	W(3,1)	W(3,2)	W(3,3)	W(3,4)

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

W

		0	1	2	3	4
A	**0**	3.12824	0.394089	0.543572	0.293635	0.733771
C	**1**	0.43404	0.284069	3.58019	0.562085	2.21896
G	**2**	0.795203	2.66042	0.593614	1.91294	0.25387
T	**3**	0.642521	1.66142	0.28262	2.23134	1.7934

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

$$W_{BASE,\ start} = \frac{\boxed{W'_{BASE,\ start}}}{\sum_{i=A,C,G,T} W'_{i,0}}$$

W

		0	1	2	3	4
A	**0**	3.12824	0.394089	0.543572	0.293635	0.733771
C	**1**	0.43404	0.284069	3.58019	0.562085	2.21896
G	**2**	0.795203	2.66042	0.593614	1.91294	0.25387
T	**3**	0.642521	1.66142	0.28262	2.23134	1.7934

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

$$W_{BASE,\ start} = \frac{W'_{BASE,\ start}}{\sum_{i=A,C,G,T} W'_{i,0}} = \frac{3.12824}{3.12824 + 0.43404 + 0.795203 + 0.642521} = 0.625647$$

W

		0	1	2	3	4
A	0	3.12824	0.394089	0.543572	0.293635	0.733771
C	1	0.43404	0.284069	3.58019	0.562085	2.21896
G	2	0.795203	2.66042	0.593614	1.91294	0.25387
T	3	0.642521	1.66142	0.28262	2.23134	1.7934

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

$$W_{BASE,\ start} = \frac{W'_{BASE,\ start}}{\sum_{i=A,C,G,T} W'_{i,0}} = \frac{3.12824}{3.12824 + 0.43404 + 0.795203 + 0.642521} = 0.625647$$

W

		0	1	2	3	4
A	0	0.625647	0.394089	0.543572	0.293635	0.733771
C	1	0.43404	0.284069	3.58019	0.562085	2.21896
G	2	0.795203	2.66042	0.593614	1.91294	0.25387
T	3	0.642521	1.66142	0.28262	2.23134	1.7934

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

W

		0	1	2	3	4
A	0	0.625647	0.0788178	0.108714	0.0587271	0.146754
C	1	0.086808	0.0568139	0.716039	0.112417	0.443791
G	2	0.159041	0.532084	0.118723	0.382587	0.050774
T	3	0.128504	0.332284	0.0565241	0.446269	0.358681

Example

Step 2: Refinement

Refine the weight matrix given the position matrix

W

		0	1	2	3	4
A	**0**	0.625647	0.0788178	0.108714	0.0587271	0.146754
C	**1**	0.086808	0.0568139	0.716039	0.112417	0.443791
G	**2**	0.159041	0.532084	0.118723	0.382587	0.050774
T	**3**	0.128504	0.332284	0.0565241	0.446269	0.358681

posM

	0	1	2	3	4	5
0	0.00132979	0.0598404	0.00132979	0.0359043	0.00398936	0.897606
1	0.163438	0.00121065	0.817191	0.00363196	0.00363196	0.0108959
2	0.25	0.0833333	0.416667	0.0277778	0.0833333	0.138889

Example

Step 2: Refinement

Refine the position matrix given the new weight matrix

W

		0	1	2	3	4
A	0					
C	1					
G	2					
T	3					

posM

	0	1	2	3	4	5
0						
1						
2						

Example

Step 2: Refinement

Refine the weight matrix given the new position matrix

W

		0	1	2	3	4
A	**0**					
C	**1**					
G	**2**					
T	**3**					

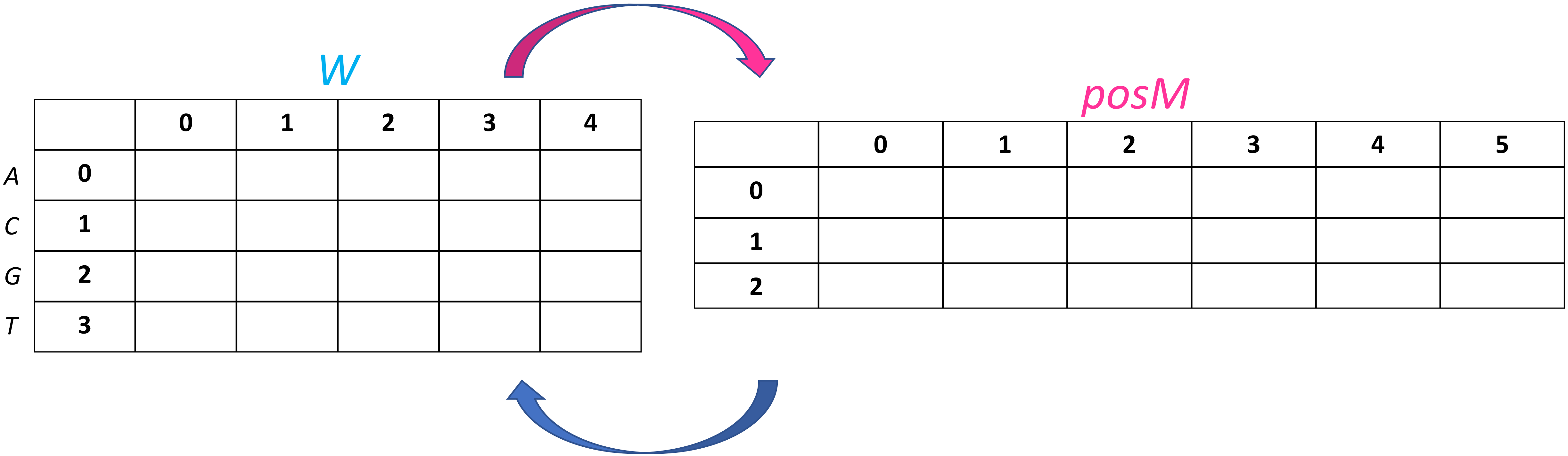
posM

	0	1	2	3	4	5
0						
1						
2						

Example

Step 2: Refinement

Refine the position matrix given the new weight matrix



W

		0	1	2	3	4
A	0					
C	1					
G	2					
T	3					

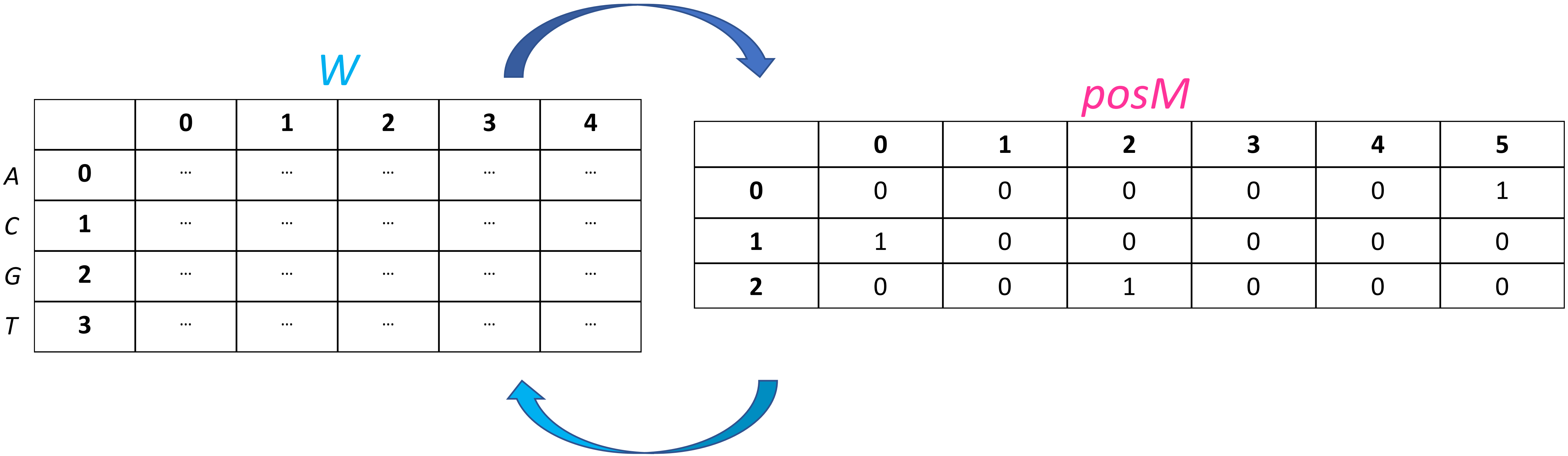
posM

	0	1	2	3	4	5
0						
1						
2						

Example

Step 2: Refinement

Until convergence



W

		0	**1**	**2**	**3**	**4**
A	**0**	…	…	…	…	…
C	**1**	…	…	…	…	…
G	**2**	…	…	…	…	…
T	**3**	…	…	…	…	…

posM

	0	**1**	**2**	**3**	**4**	**5**
0	0	0	0	0	0	1
1	1	0	0	0	0	0
2	0	0	1	0	0	0

Example

Step 3: Consensus sequence

sequences

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	A	G	C	T	C
1	A	G	A	T	C	G	T	A	A	C
2	T	G	G	G	C	T	A	C	C	T

posM

	0	**1**	**2**	**3**	**4**	**5**
0	0	0	0	0	0	1
1	1	0	0	0	0	0
2	0	0	1	0	0	0

Use the position matrix to find the most probable starting positions of the motif in each sequence.

Example

Step 3: Consensus sequence

sequences

	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	**A**	G	C	T	C
1	**A**	G	A	T	C	G	T	A	A	C
2	T	G	**G**	G	C	T	A	C	C	T

posM

	0	**1**	**2**	**3**	**4**	**5**
0	0	0	0	0	0	**1**
1	**1**	0	0	0	0	0
2	0	0	**1**	0	0	0

Use the position matrix to find the l-mer which is the best candidate for the motif in each sequence.

Example

Step 3: Consensus sequence

sequences

	0	1	2	3	4	5	6	7	8	9	
0	C	C	G	C	G	**A**	G	C	T	C	→ AGCTC
1	**A**	G	A	T	C	G	T	A	A	C	→ AGATC
2	T	G	**G**	G	C	T	A	C	C	T	→ GGCTA

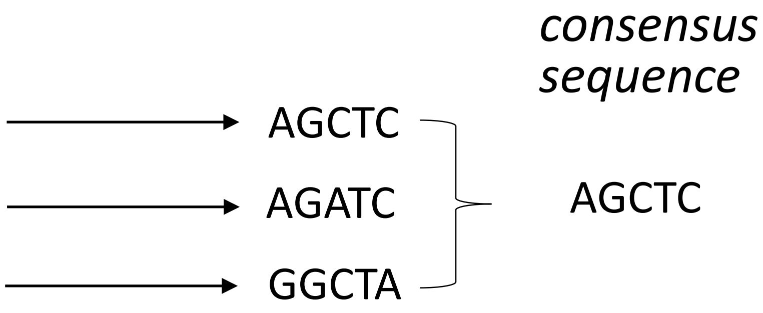
Use the position matrix to find the l-mer which is the best candidate for the motif in each sequence.

Example

Step 3: Consensus sequence

sequences

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	**A**	G	C	T	C
1	**A**	G	A	T	C	G	T	A	A	C
2	T	G	**G**	G	C	T	A	C	C	T



AGCTC

AGATC

GGCTA

consensus sequence

AGCTC

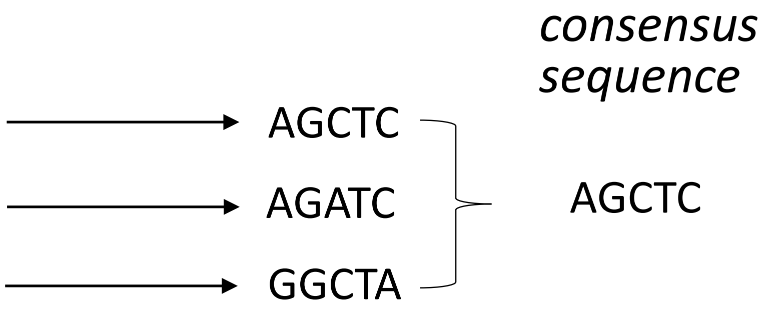
Use the position matrix to find the l-mer which is the best candidate for the motif in each sequence.

Example

Step 3: Consensus sequence

sequences

	0	1	2	3	4	5	6	7	8	9
0	C	C	G	C	G	**A**	G	C	T	C
1	**A**	G	A	T	C	G	T	A	A	C
2	T	G	**G**	G	C	T	A	C	C	T



AGCTC

AGATC

GGCTA

consensus sequence

AGCTC

hammDist(AGCTC, AGCTC) = 0

hammDist(AGCTC, AGATC) = 1

hammDist(AGCTC, GGCTA) = 2

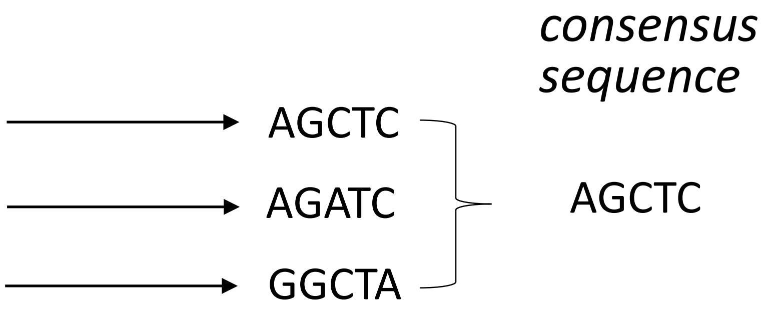
Calculate the score of the bucket: the number of l-mers whose hamming distance to the consensus sequence exeeds the maximum number of mutations d.

Example

Step 3: Consensus sequence

sequences

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	**A**	G	C	T	C
1	**A**	G	A	T	C	G	T	A	A	C
2	T	G	**G**	G	C	T	A	C	C	T



consensus sequence

AGCTC

AGATC

GGCTA

AGCTC

hammDist(AGCTC, AGCTC) = 0

hammDist(AGCTC, AGATC) = 1

hammDist(AGCTC, GGCTA) = 2

Calculate the score of the bucket: the number of l-mers whose hamming distance to the consensus sequence exeeds the maximum number of mutations d.

d = 2

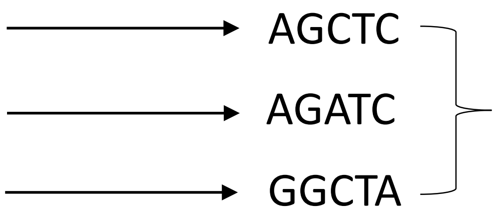
the score of this bucket is 0

Example

Step 3: Consensus sequence

sequences

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	C	C	G	C	G	**A**	G	C	T	C
1	**A**	G	A	T	C	G	T	A	A	C
2	T	G	**G**	G	C	T	A	C	C	T



AGCTC

AGATC

GGCTA

consensus sequence

AGCTC

hammDist(AGCTC, AGCTC) = 0

hammDist(AGCTC, AGATC) = 1

hammDist(AGCTC, GGCTA) = 2

For each bucket save the score and the consensus sequence.

d = 2

the score of this bucket is 0

Example

Step 3: Consensus sequence

bucket0 - *consSeq0*, *score0*
bucket1 - *consSeq1*, *score1*
.
.
.
bucketN - *consSeqN*, *scoreN*

For each bucket save the score and the consensus sequence.

Example

Step 3: Consensus sequence

bucket0 - *consSeq0*, *score0*

bucket1 - *consSeq1*, *score1*

.

.

.

bucketN - *consSeqN*, *scoreN*

bucketX - *consSeqX*, *scoreX*

Keep the bucket with the best score scoreX

Example

Take the best consensus sequence from all trials

„AGCTC“

trial 1

bucketX - *consSeqX*, *scoreX*

trial 2

bucketX - *consSeqX*, *scoreX*

• • •

trial m

bucketX - *consSeqX*, *scoreX*

Outline

Introduction & Motivation

Problem

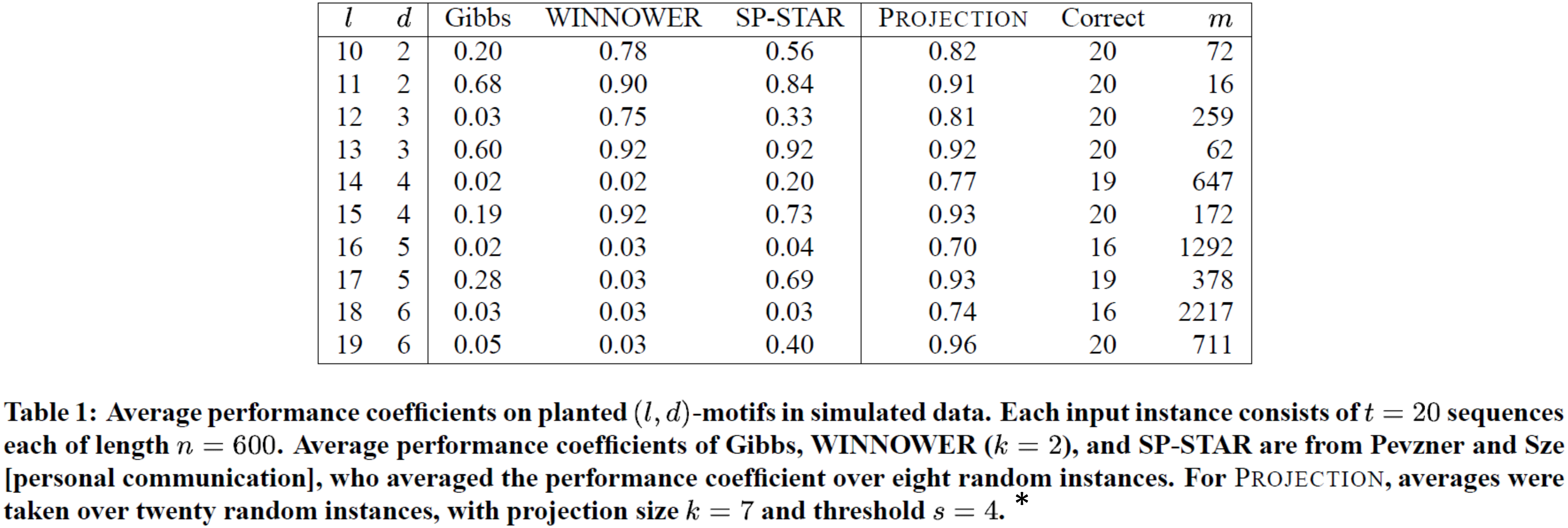
Background

Solution

➢ Validation

Future work

Validation

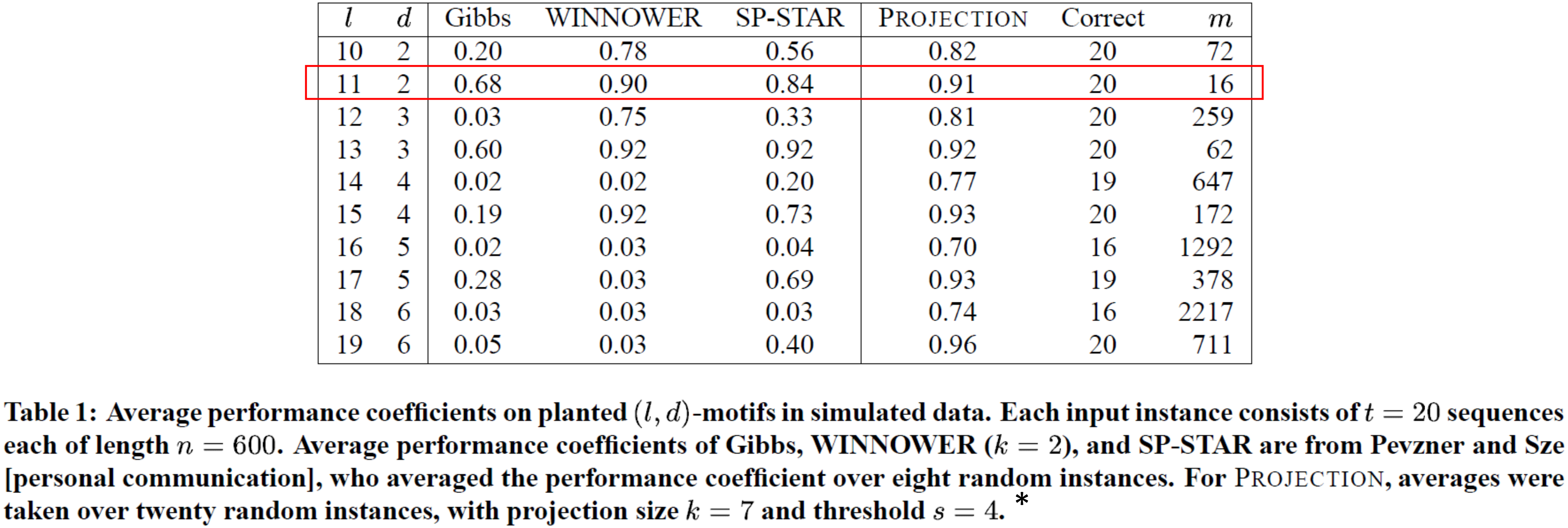


l	d	Gibbs	WINNOWER	SP-STAR	PROJECTION	Correct	m
10	2	0.20	0.78	0.56	0.82	20	72
11	2	0.68	0.90	0.84	0.91	20	16
12	3	0.03	0.75	0.33	0.81	20	259
13	3	0.60	0.92	0.92	0.92	20	62
14	4	0.02	0.02	0.20	0.77	19	647
15	4	0.19	0.92	0.73	0.93	20	172
16	5	0.02	0.03	0.04	0.70	16	1292
17	5	0.28	0.03	0.69	0.93	19	378
18	6	0.03	0.03	0.03	0.74	16	2217
19	6	0.05	0.03	0.40	0.96	20	711

Table 1: Average performance coefficients on planted (l, d)-motifs in simulated data. Each input instance consists of $t = 20$ sequences each of length $n = 600$. Average performance coefficients of Gibbs, WINNOWER ($k = 2$), and SP-STAR are from Pevzner and Sze [personal communication], who averaged the performance coefficient over eight random instances. For PROJECTION, averages were taken over twenty random instances, with projection size $k = 7$ and threshold $s = 4$. *

* Jeremy Buhler and Martin Tompa. Finding motifs using random projections. J Comput Biol. 2002. 9(2):225-42

Validation



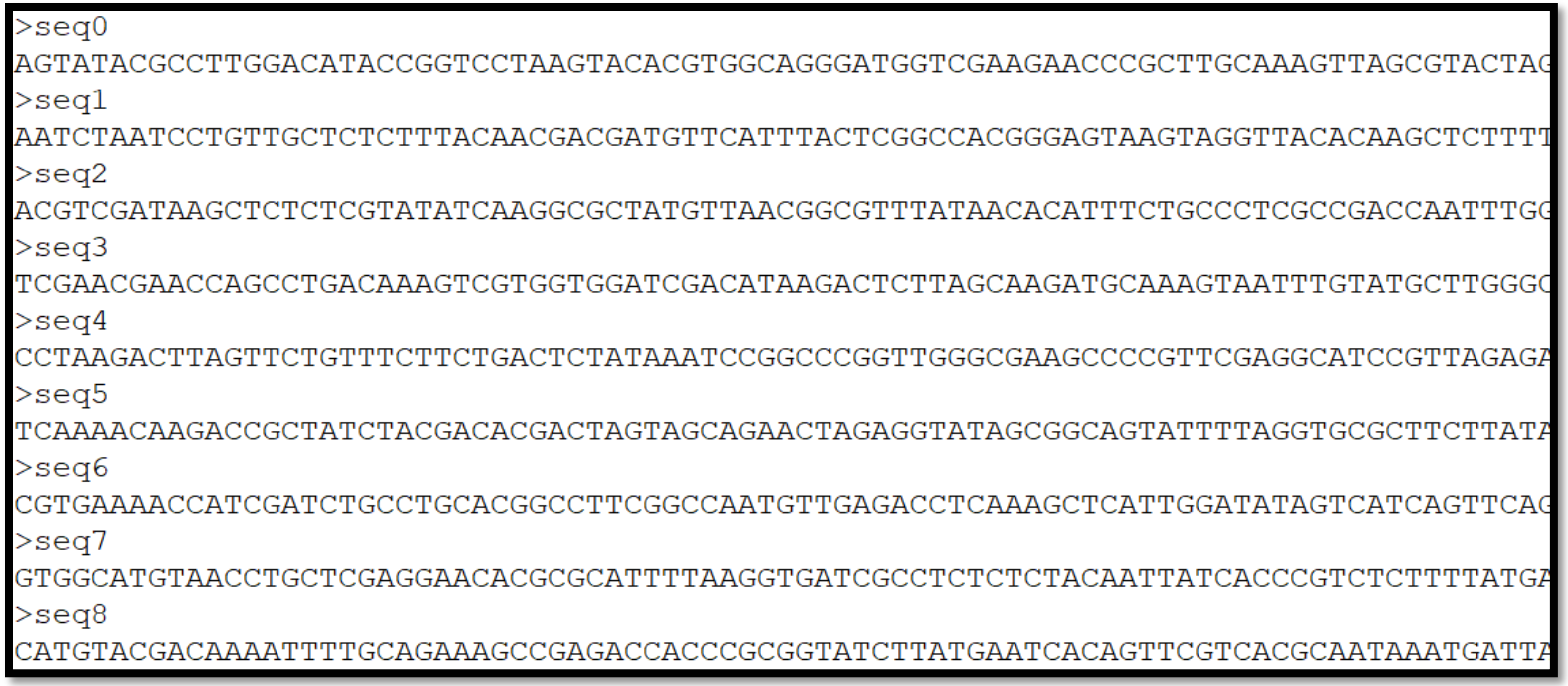
l	d	Gibbs	WINNOWER	SP-STAR	PROJECTION	Correct	m
10	2	0.20	0.78	0.56	0.82	20	72
11	2	0.68	0.90	0.84	0.91	20	16
12	3	0.03	0.75	0.33	0.81	20	259
13	3	0.60	0.92	0.92	0.92	20	62
14	4	0.02	0.02	0.20	0.77	19	647
15	4	0.19	0.92	0.73	0.93	20	172
16	5	0.02	0.03	0.04	0.70	16	1292
17	5	0.28	0.03	0.69	0.93	19	378
18	6	0.03	0.03	0.03	0.74	16	2217
19	6	0.05	0.03	0.40	0.96	20	711

Table 1: Average performance coefficients on planted (l, d)-motifs in simulated data. Each input instance consists of $t = 20$ sequences each of length $n = 600$. Average performance coefficients of Gibbs, WINNOWER ($k = 2$), and SP-STAR are from Pevzner and Sze [personal communication], who averaged the performance coefficient over eight random instances. For PROJECTION, averages were taken over twenty random instances, with projection size $k = 7$ and threshold $s = 4$. *

16 trials for 20 sequences of length 600 with a planted motif AGGCATCCGTT of length 11 with max 2 mutations, k-mer projection size 7 and bucket threshold 4.

* Jeremy Buhler and Martin Tompa. Finding motifs using random projections. J Comput Biol. 2002. 9(2):225-42

Validation



```
>seq0
AGTATACGCCTTGGACATACCGGTCCTAAGTACACGTGGCAGGGATGGTCGAAGAACCCGCTTGCAAAGTTAGCGTACTAG
>seq1
AATCTAATCCTGTTGCTCTCTTTACAACGACGATGTTCATTTACTCGGCCACGGGAGTAAGTAGGTTACACAAGCTCTTTT
>seq2
ACGTCGATAAGCTCTCTCGTATATCAAGGCGCTATGTTAACGGCGTTTATAACACATTTCTGCCCTCGCCGACCAATTTGG
>seq3
TCGAACGAACCAGCCTGACAAAGTCGTGGTGGATCGACATAAGACTCTTAGCAAGATGCAAAGTAATTTGTATGCTTGGGC
>seq4
CCTAAGACTTAGTTCTGTTTCTTCTGACTCTATAAATCCGGCCCGGTTGGGCGAAGCCCCGTTCGAGGCATCCGTTAGAGA
>seq5
TCAAAACAAGACCGCTATCTACGACACGACTAGTAGCAGAACTAGAGGTATAGCGGCAGTATTTTAGGTGCGCTTCTTATA
>seq6
CGTGAAAACCATCGATCTGCCTGCACGGCCTTCGGCCAATGTTGAGACCTCAAAGCTCATTGGATATAGTCATCAGTTCAG
>seq7
GTGGCATGTAACCTGCTCGAGGAACACGCGCATTTTAAGGTGATCGCCTCTCTCTACAATTATCACCCGTCTCTTTTATGA
>seq8
CATGTACGACAAAATTTTGCAGAAAGCCGAGACCACCCGCGGTATCTTATGAATCACAGTTCGTCACGCAATAAATGATTA
```

16 trials for 20 sequences of length 600 with a planted motif AGGCATCCGTT of length 11 with max 2 mutations, k-mer projection size 7 and bucket threshold 4.

Validation

Planted motif: `AGGCATCCGTT`

Motif finder	Time
PROJECTION	1h 30min
MEME	15.74 secs
Gibbs sampler	47.81 secs

Outline

Introduction & Motivation

Problem

Background

Solution

Validation

➢Future work

Future work

- 0 or more than 1 occurences per sequence (ZOOPS, TCM)
- Multiple cores
- Unknown DNA-base N (SEQAN Dna5)
- Hash buckets of SEQAN
- Tests with more datasets