BIOINFORMATICS

How do we locate disease causing mutation?

Marco Beccuti

Università degli Studi di Torino Dipartimento di Informatica



Exercise 1 Finding the Matched Patterns

Exercise

- For reference bioinformatics create:
 - ► BWT;
 - FirstOccurence;
 - CheckPointArray_b with b=5;
 - SuffixArray_a with a=5;
 - ► Search Pattern: ioi

3/13

- For reference bioinformatics create:
 - ► BWT.

Text: bioinformatics\$

Rotations: bioinformatics\$ ioinformatics\$b oinformatics\$bi informatics\$bio nformatics\$bioi formatics\$bioin ormatics\$bioinf rmatics\$bioinfo matics\$bioinfor atics\$bioinform tics\$bioinforma ics\$bioinformat cs\$bioinformati s\$bioinformatic \$bioinformatics

- For reference **bioinformatics** create:
 - BWT.

Text:

bioinformatics\$

Rotations: bioinformaticss ioinformatics\$b oinformatics\$bi informatics\$bio nformatics\$bioi formatics\$bioin ormatics\$bioinf rmatics\$bioinfo matics\$bioinfor atics\$bioinform tics\$bioinforma ics\$bioinformat cs\$bioinformati s\$bioinformatic \$bioinformatics

Sorted Rotations: **\$bioinformatics** atics\$bioinform bioinformatics\$ cs\$bioinformati formatics\$bioin ics\$bioinformat informatics\$bio ioinformatics\$b matics\$bioinfor nformatics\$bioi oinformatics\$bi ormatics\$bioinf rmatics\$bioinfo s\$bioinformatic tics\$bioinforma

- For reference bioinformatics create:
 - BWT.

Text: bioinformatics\$

```
BWT matrix
        bioinformatic
$
                          S
        tics$bioinfor
a
                          m
b
        ioinformatics
                          $
        s$bioinformat
C
f
        ormatics$bioi
                          n
i
        cs$bioinforma
                          t
i
        nformatics$bi
                          0
i
        oinformatics$
                          b
        atics$bioinfo
m
                          r
        formatics$bio
n
        informatics$b
                          i
0
        rmatics$bioin
0
        matics$bioinf
r
                          0
        $bioinformati
s
                          C
        ics$bioinform
t
                          a
```

- For reference bioinformatics create:
 - ► FirstOccurence.

Text: bioinformatics\$

| BWT r | matrix | | FirstOccurence |
|-------|----------------|----|----------------|
| \$ | bioinformatic | s | 0 |
| a | tics\$bioinfor | m | 1 |
| b | ioinformatics | \$ | 2 |
| C | s\$bioinformat | i | 3 |
| f | ormatics\$bioi | n | 4 |
| i | cs\$bioinforma | t | 5 |
| i | nformatics\$bi | 0 | |
| i | oinformatics\$ | b | |
| m | atics\$bioinfo | r | 8 |
| n | formatics\$bio | i | 9 |
| 0 | informatics\$b | i | 10 |
| 0 | rmatics\$bioin | f | |
| r | matics\$bioinf | 0 | 12 |
| S | \$bioinformati | С | 13 |
| t | ics\$bioinform | a | 14 |

M. Beccuti BIOINFORMATICS 7/13

- For reference bioinformatics create:
 - ► CheckPointArray_b with b=5.

Text: bioinformatics\$ a=b=5

| BWT m | atrix | | FirstOccurence |
|-------|----------------|----|----------------|
| \$ | bioinformatic | s< | 0 |
| a | tics\$bioinfor | m | 1 |
| b | ioinformatics | \$ | 2 |
| С | s\$bioinformat | i | 3 |
| f | ormatics\$bioi | n | 4 |
| i | cs\$bioinforma | t< | 5 |
| i | nformatics\$bi | 0 | |
| i | oinformatics\$ | b | |
| m | atics\$bioinfo | r | 8 |
| n | formatics\$bio | i | 9 |
| 0 | informatics\$b | i< | 10 |
| 0 | rmatics\$bioin | f | |
| r | matics\$bioinf | 0 | 12 |
| S | \$bioinformati | С | 13 |
| t | ics\$bioinform | a | 14 |

• For reference **bioinformatics** create:

► Suffix Array_a with a=5.

Text: bioinformatics\$ a=b=5

| BWT | matrix | | FirstOccurence | Ś | а | b | С | f | i | m | n | 0 | r | s | t | Suffix array |
|-----|----------------|----|----------------|---|---|---|-----|---|---|---|---|---|---|-----|---|--------------|
| \$ | bioinformatic | s< | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | , |
| a | tics\$bioinfor | m | 1 | | | | | | | | | | | | | |
| b | ioinformatics | \$ | 2 | | | | | | | | | | | | | 0 |
| C | s\$bioinformat | i | 3 | | | | | | | | | | | | | |
| f | ormatics\$bioi | n | 4 | | | | | | | | | | | | | 5 |
| i | cs\$bioinforma | t< | 5 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | |
| i | nformatics\$bi | 0 | | | | | | | | | | | | | | |
| i | oinformatics\$ | b | | | | | | | | | | | | | | |
| m | atics\$bioinfo | r | 8 | | | | | | | | | | | | | |
| n | formatics\$bio | i | 9 | | | | | | | | | | | | | |
| 0 | informatics\$b | i< | 10 | 1 | 0 | 1 | . 0 | 0 | 2 | 1 | 1 | 1 | 1 | . 1 | 1 | |
| 0 | rmatics\$bioin | f | | | | | | | | | | | | | | |
| r | matics\$bioinf | 0 | 12 | | | | | | | | | | | | | |
| s | \$bioinformati | С | 13 | | | | | | | | | | | | | |
| t | ics\$bioinform | a | 14 | | | | | | | | | | | | | 10 |

M. Beccuti 9/13

Text: bioinformatics\$

-->

For reference bioinformatics create:

Search Pattern: ioi

```
a=b=5
Search =ioi
Top=0 Bottom=14
        BWT matrix
                                         FirstOccurence
                bioinformatic
                                                 0
-->
                tics$bioinfor
                                                 1
        b
                ioinformatics
                s$bioinformat
                ormatics$bioi
                cs$bioinforma
                                 t
                nformatics$bi
                oinformatics$
                atics$bioinfo
                                                 8
                formatics$bio
                informaticsSb
                                                 10
                rmatics$bioin
                matics$bioinf
                                                 12
                $bioinformati
                                                 13
                ics$bioinform
```

```
$ a b c f i m n o r s t
                          Suffix array
0 0 0 0 0 0 0 0 0 0 0
                               0
                               5
100001110010
101002111111
                              10
```

Top= FirstOccurance(i)+Count i(0,sm\$intobriifoca)=5 Bottom= FirstOccurance(i)+Count_i(15,sm\$intobriifoca)-1=7

14

• For reference bioinformatics create:

► Search Pattern: ioi

Text: bioinformatics\$
a=b=5
Search =jo->i
Top=5 Bottom=7

| | <u>BWT</u> m | natrix | | FirstOccurence | \$ | а | b | С | f | i | m | n | 0 | r | s | t | Suf | fix ar | ray |
|---|--------------|------------------------|----|----------------|----|---|---|---|---|---|---|---|---|---|---|---|-----|--------|-----|
| | \$ | bioinformatic | s | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | | |
| | a | tics\$bioinfor | m | 1 | | | | | | | | | | | | | | | |
| | b | ioinformatics | \$ | 2 | | | | | | | | | | | | | | Θ | |
| | С | s\$ <u>bioinformat</u> | i | 3 | | | | | | | | | | | | | | | |
| | f | ormatics\$bioi | n | 4 | | | | | | | | | | | | | | 5 | |
| > | i | cs\$ <u>bioinforma</u> | t | 5 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | | | |
| | i | nformatics\$bi | 0 | | | | | | | | | | | | | | | | |
| > | i | oinformatics\$ | b | | | | | | | | | | | | | | | | |
| | m | atics\$bioinfo | r | 8 | | | | | | | | | | | | | | | |
| | n | formatics\$bio | i | 9 | | | | | | | | | | | | | | | |
| | 0 | informatics\$b | i | 10 | 1 | 0 | 1 | 0 | 0 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | | | |
| | 0 | rmatics\$bioin | f | | | | | | | | | | | | | | | | |
| | r | matics\$bioinf | 0 | 12 | | | | | | | | | | | | | | | |
| | s | \$bioinformati | С | 13 | | | | | | | | | | | | | | | |
| | t | ics\$bioinform | а | 14 | | | | | | | | | | | | | | 10 | |

Top= FirstOccurance(o)+Count_o(5,sm\$intobriifoca)=10
Bottom= FirstOccurance(o)+Count_o(8,sm\$intobriifoca)-1=10

• For reference bioinformatics create:

► Search Pattern: ioi

Text: bioinformatics\$
a=b=5
Search =i->oi
Top=10 Bottom=10

| | BWT mat | trix | | FirstOccurence | \$ | i a | a b |) с | f | i | m | n | 0 | r | s | t | Suffi | x array |
|---|---------|------------------------|----|----------------|----|-----|-----|-----|---|---|---|---|---|---|---|---|-------|---------|
| | \$ | bioinformatic | s | 0 | Θ |) (| 9 (| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | |
| | a | tics\$ <u>bioinfor</u> | m | 1 | | | | | | | | | | | | | | |
| | b | ioinformatics | \$ | 2 | | | | | | | | | | | | | | 0 |
| | С | s\$bioinformat | i | 3 | | | | | | | | | | | | | | |
| | f | ormatics\$bioi | n | 4 | | | | | | | | | | | | | | 5 |
| | i | cs\$ <u>bioinforma</u> | t | 5 | 1 | . 6 | 9 6 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | | |
| | i | nformatics\$bi | 0 | | | | | | | | | | | | | | | |
| | i | oinformatics\$ | b | | | | | | | | | | | | | | | |
| | m | atics\$bioinfo | r | 8 | | | | | | | | | | | | | | |
| | n | formatics\$bio | i | 9 | | | | | | | | | | | | | | |
| > | 0 | informatics\$b | i | 10 | 1 | . 6 |)] | . 0 | 0 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | | |
| | 0 | rmatics\$bioin | f | | | | | | | | | | | | | | | |
| | r | matics\$bioinf | 0 | 12 | | | | | | | | | | | | | | |
| | s | \$bioinformati | С | 13 | | | | | | | | | | | | | | |
| | t | ics\$bioinform | а | 14 | | | | | | | | | | | | | | 10 |

Top= FirstOccurance(i)+Count_i(10,sm\$intobriifoca)=7
Bottom= FirstOccurance(i)+Count_i(11,sm\$intobriifoca)-1=7

• For reference bioinformatics create:

► Search Pattern: ioi

Text: bioinformatics\$
a=b=5
Search =->ioi
Top=7 Bottom=7

-->

| BWT ma | ntrix | | FirstOccurence | \$ | а | b | С | f | i | m | n | 0 | r | s | t |
|--------|------------------------|----|----------------|----|---|---|---|---|---|---|---|---|---|---|---|
| \$ | bioinformatic | s | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| а | tics\$bioinfor | m | 1 | | | | | | | | | | | | |
| b | ioinformatics | \$ | 2 | | | | | | | | | | | | |
| С | s\$ <u>bioinformat</u> | i | 3 | | | | | | | | | | | | |
| f | ormatics\$bioi | n | 4 | | | | | | | | | | | | |
| i | cs\$ <u>bioinforma</u> | t | 5 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 |
| i | nformatics\$bi | 0 | | | | | | | | | | | | | |
| i | oinformatics\$ | b | | | | | | | | | | | | | |
| m | atics\$bioinfo | r | 8 | | | | | | | | | | | | |
| n | formatics\$bio | i | 9 | | | | | | | | | | | | |
| 0 | informatics\$b | i | 10 | 1 | 0 | 1 | 0 | 0 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0 | rmatics\$bioin | f | | | | | | | | | | | | | |
| r | matics\$bioinf | 0 | 12 | | | | | | | | | | | | |
| s | \$bioinformati | С | 13 | | | | | | | | | | | | |
| t | ics\$bioinform | a | 14 | | | | | | | | | | | | |

M. Beccuti BIOINFORMATICS 13/13

Suffix array

5