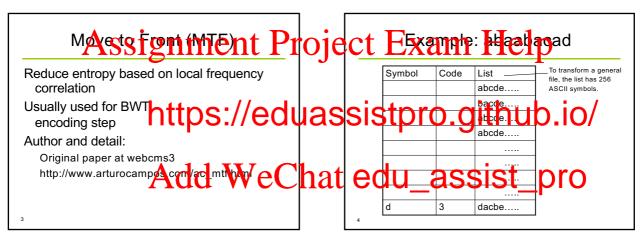
## COMP9319 Web Data Compression and Search

BWT, MTF and Pattern Matching

### **BWT**

- Burrows-Wheeler transform (BWT) is an algorithm used to prepare data for use with data compression techniques such as bzip2.
- It was invented by Michael Burrows and David Wheeler in 1994 at DEC SRC, Palo Alto,
- It is based on a previously unpublished transformation discovered by Wheeler in 1983.

1



### Example: abaaabbbccddddcc

Symbols: abaaabbbccddddcc

Codes (in ASCII binary): 01100001, 01100010, 01100001, 01100001, ...,

01100100, 01100011, 01100011 Codes (in ASCII dec): 97, 98, 97, 97, 97, 98, 98, 98, 99, 99, 100, 100, 100,

100, 99, 99

### Example: abaaabbbccddddcc

6

Symbols: abaaabbbccddddcc Codes (in ASCII binary): 01100001, 01100010, 01100001, 01100001, ...,

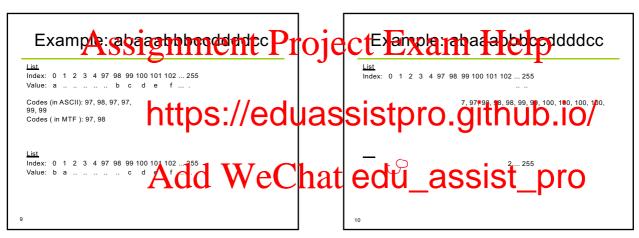
01100100, 01100011, 01100011 Codes (in ASCII dec): 97, 98, 97, 97, 97, 98, 98, 98, 99, 99, 100, 100, 100, 100, 99, 99

> Recall that Shannon's entropy reaches the max when there is max uncertainly, i.e., equal probability, like the example above (4 "97"s, 4 "98"s, 4 "99"s, 4 "100"s).

e.g., Entropy H = 2.00

# Example: abaaabbbccddddcc Codes (in ASCII binary): 01100010, 01100001, 01100001, 01100001, ..., 01100100, 01100011, 01100001 Codes (in ASCII dec): 97, 98, 97, 97, 98, 98, 98, 99, 99, 100, 100, 100, 100, 99, 99 List Index: 0 1 2 3 4 97 98 99 100 101 102 ... 255 Value: ... ... ... ... a b c d e f ... ... Codes (in ASCII): 97, 98, 97, 97, 98, 98, 98, 99, 99, 100, 100, 100, 100, 99, 99 Codes (in MTF): 97

7



8

9 10

```
Example: abaaabbbccddddcc

List.
Index: 0 1 2 3 4 97 98 99 100 101 102 ... 255
Value: a b ... ... ... c d e f ... ...

Codes (in ASCII): 97, 98, 97, 97, 97, 98, 98, 98, 99, 99, 100, 100, 100, 100, 99, 99
Codes ( in MTF ): 97, 98, 1, 0,

List.
Index: 0 1 2 3 4 97 98 99 100 101 102 ... 255
Value: a b ... ... ... c d e f ... ...
```

```
List Index: 0 1 2 3 4 97 98 99 100 101 102 ... 255 Value: a b ... ... ... ... c d e f ... ...

Codes (in ASCII): 97, 98, 97, 97, 97, 98, 98, 99, 99, 100, 100, 100, 100, 99, 99

Codes (in MTF): 97, 98, 1, 0, 0, 1, 0, 0, 99, 0, 100, 0, 0, 0, 1, 0
```

11 12

## List Index: 0 1 2 3 4 97 98 99 100 101 102 ... 255 Value: ... ... ... a b c d e f ... ... Codes (in MTF): 97, 98, 1, 0, 0, 1, 0, 0, 99, 0, 100, 0, 0, 0, 1, 0 Symbols: a, b

```
Example: MTF decoding

List
Index: 0 1 2 3 4 97 98 99 100 101 102 ... 255
Value: b a ... ... ... c d e f ... ..

Codes (in MTF): 97, 98, 1, 0, 0, 1, 0, 0, 99, 0, 100, 0, 0, 0, 1, 0
Symbols: a, b, a

List
Index: 0 1 2 3 4 97 98 99 100 101 102 ... 255
Value: a b ... ... ... c d e f ... ..
```

13 14

```
Examples METAPOINT Project Examples METAPOINT Pr
```

15 16

```
Example: MTF decoding

List
Index: 0 1 2 3 4 97 98 99 100 101 102 ... 255
Value: a b ... ... ... ... c d e f ... ..

Codes (in MTF): 97, 98, 1, 0, 0, 1, 0, 0, 99, 0, 100, 0, 0, 0, 1, 0
Symbols: a, b, a, a, a, b, b, b, c, c, d, d, d, d, c, c

The distribution of symbols is changed, with more local references (1 "97", 1 "98", 1 "99", 1 "100", 9
"0"s, 3 "1"s). => Reduced entropy

H = 1.92
```

ZIP (i.e	., LZW based)		BWT+RLE+MTF+AC			
File Name	Raw Size	PKZIP Size	PKZIP Bits/Byte	BWT Size	BWT Bits/Byte	
bib	111,261	35,821	2.58	29,567	2.13	
book1	768,771	315,999	3.29	275,831	2.87	
book2	610,856	209,061	2.74	186,592	2.44	
geo	102,400	68,917	5.38	62,120	4.85	
news	377,109	146,010	3.10	134,174	2.85	
obj1	21,504	10,311	3.84	10,857	4.04	
obj2	246,814	81,846	2.65	81,948	2.66	

17 18