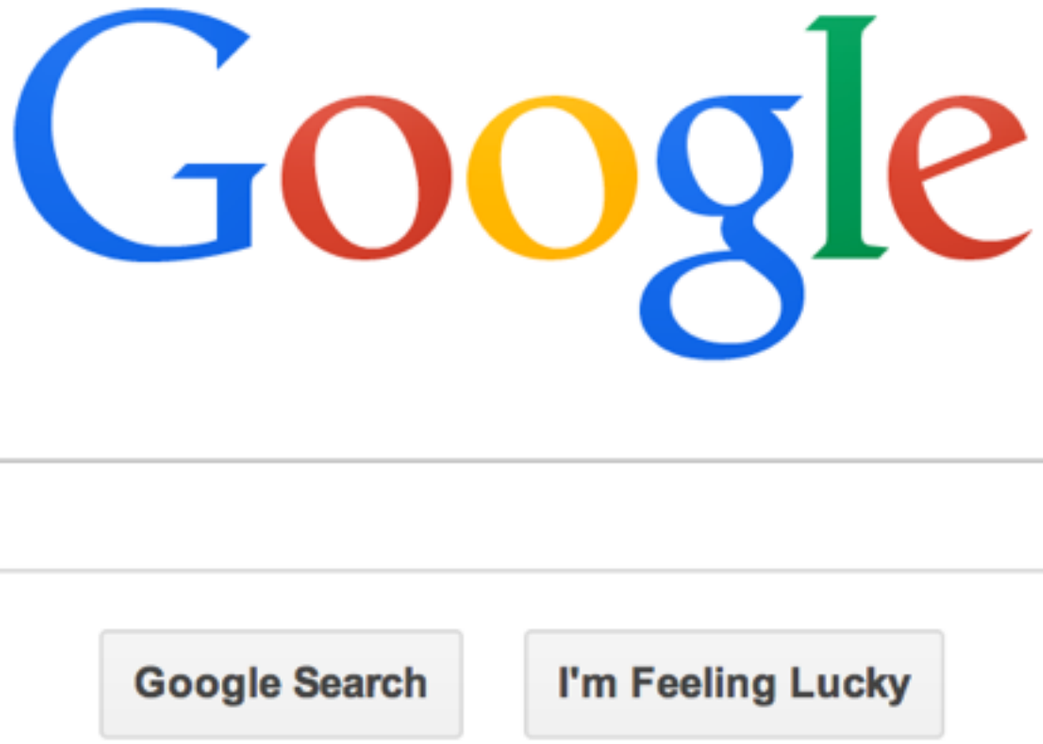
Indexing



Indexing

Index of T

```
CGTGC: 0,4
GCGTG: 3
GTGCC: 1
GTGCT: 5
TGCCT: 2
TGCTT: 6
```

T: C G T G C G T G C T T

Indexing subsequences

Index of T

C G G G T : **0**
G T C T G : **1**
T G G G C : **2**

T: C G T G C G T G C T T

PatternHunter

BIOINFORMATICS

Vol. 18 no. 3 2002
Pages 440–445

PatternHunter: faster and more sensitive homology search

Bin Ma[1], John Tromp[2] and Ming Li[3]

[1]Computer Science Department, University of Western Ontario, London N6A 5B8, Canada, [2]Bioinformatics Solutions Inc., 145 Columbia Street West, Waterloo, Ont. N2L 3L2, Canada and [3]Computer Science Department, University of Waterloo, Waterloo, Ont. N2L 3G1, Canada and Bioinformatics Lab, Computer Science Department, University of California, Santa Barbara, CA 93106, USA

Received on August 24, 2001; revised on October 10, 2001; accepted on October 15, 2001

Suffix index

T: G T T A T A G C T G A T C G C G G C G T A G C G G

G T T A T A G C T G A T C G C G G C G T A G C G G
T T A T A G C T G A T C G C G G C G T A G C G G
T A T A G C T G A T C G C G G C G T A G C G G
A T A G C T G A T C G C G G C G T A G C G G
T A G C T G A T C G C G G C G T A G C G G
A G C T G A T C G C G G C G T A G C G G
G C T G A T C G C G G C G T A G C G G
C T G A T C G C G G C G T A G C G G
T G A T C G C G G C G T A G C G G
G A T C G C G G C G T A G C G G
A T C G C G G C G T A G C G G
T C G C G G C G T A G C G G
C G C G G C G T A G C G G
G C G G C G T A G C G G
C G G C G T A G C G G
G G C G T A G C G G
G C G T A G C G G
C G T A G C G G
G T A G C G G
T A G C G G
A G C G G
G C G G
C G G
G G
G

Suffix index

T = abaaba

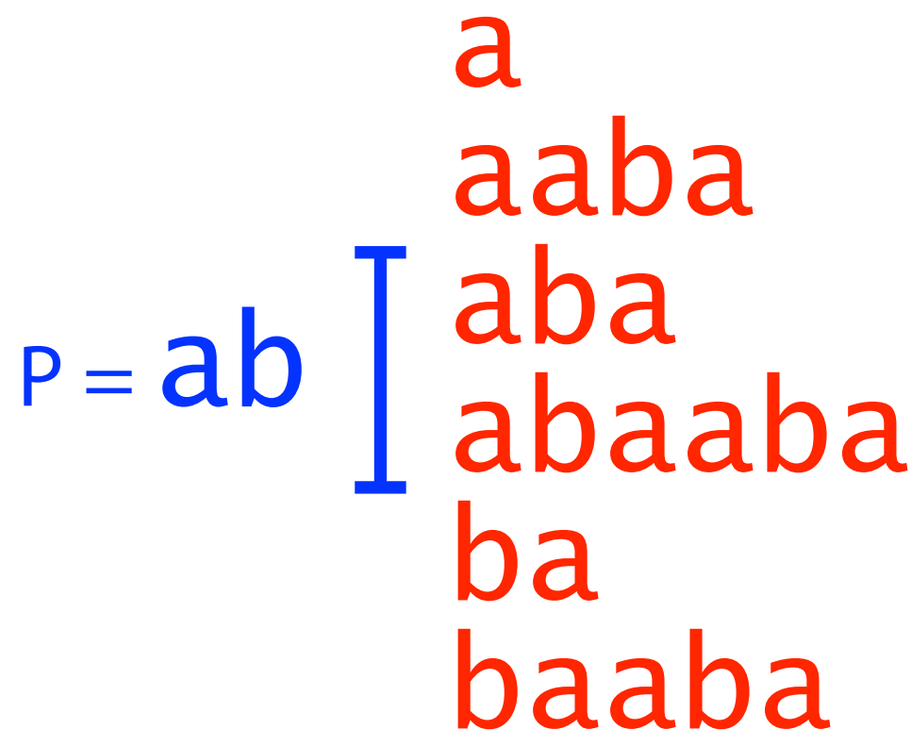
abaaba
baaba
aaba
aba
ba
a

Alphabetical order

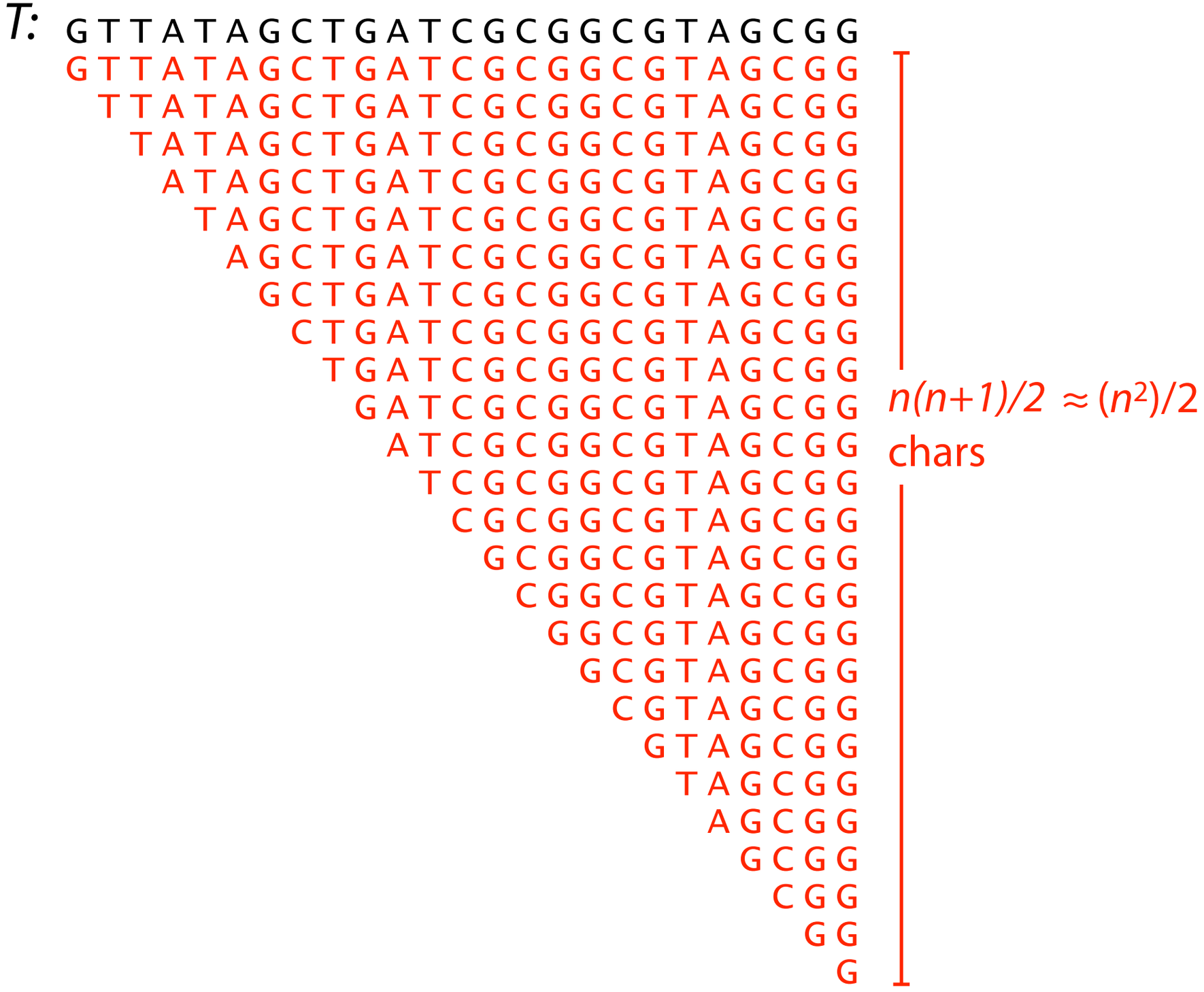
a
aaba
aba
abaaba
ba
baaba

Suffix index

Querying uses binary search



Suffix index



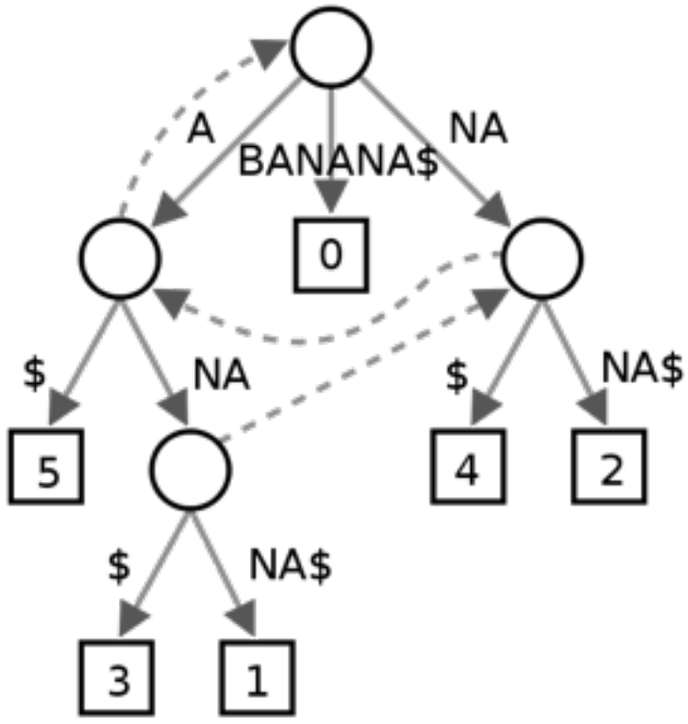
Suffix index

$T =$ abaaba

5	a
2	aaba
3	aba
0	abaaba
4	ba
1	baaba

Suffix array is m integers long

SuffixArray(T)



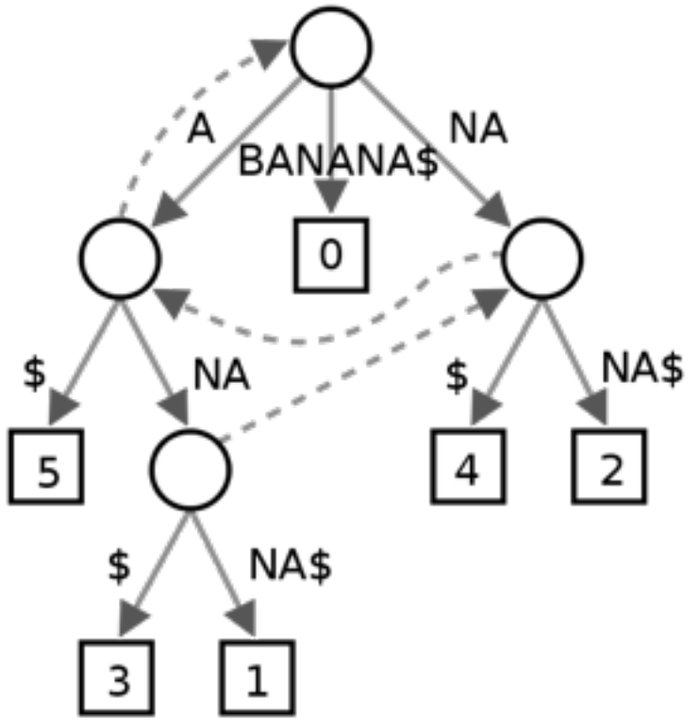
Suffix tree

6	$
5	A$
3	ANA$
1	ANANA$
0	BANANA$
4	NA$
2	NANA$

Suffix array

$ B A N A N A
A $ B A N A N
A N A $ B A N
A N A N A $ B
B A N A N A $
N A $ B A N A
N A N A $ B A

FM Index



Suffix tree

≥ 45 GB

6	$
5	A$
3	ANA$
1	ANANA$
0	BANANA$
4	NA$
2	NANA$

Suffix array

≥ 12 GB

$ B A N A N A
A $ B A N A N
A N A $ B A N
A N A N A $ B
B A N A N A $
N A $ B A N A
N A N A $ B A

FM Index

~ 1 GB