

Suffix Arrays

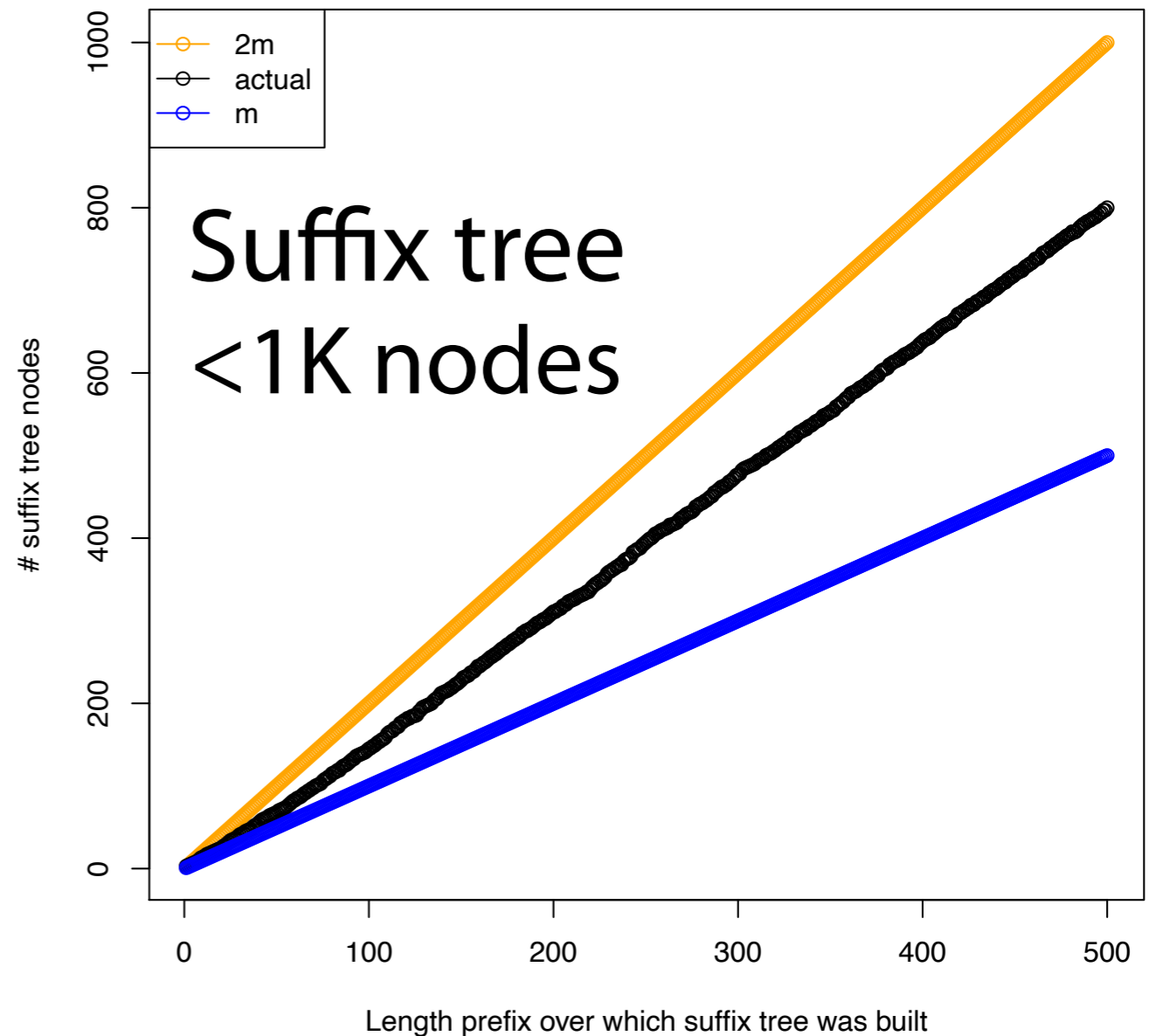
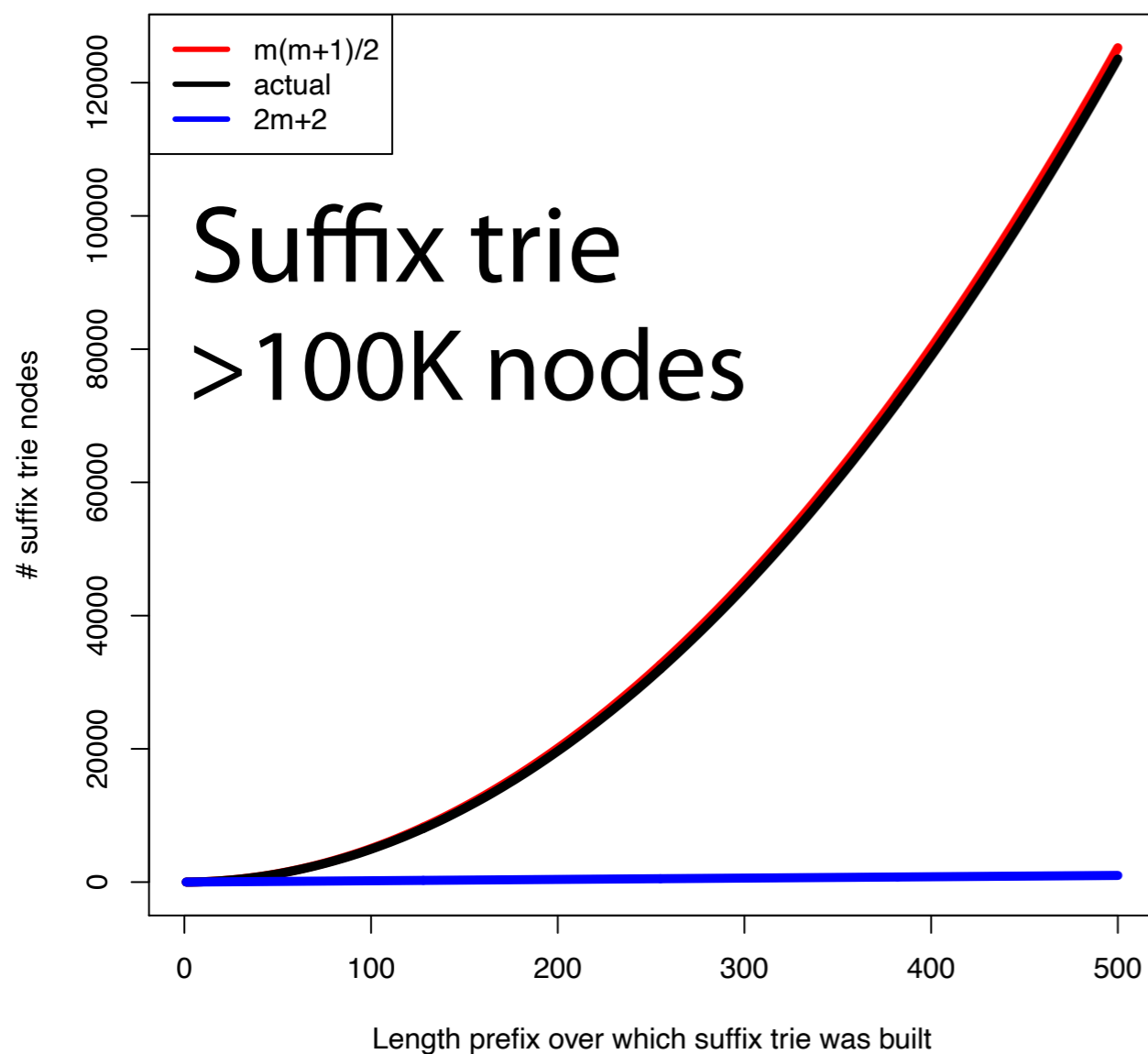
Ben Langmead



Please sign guestbook (www.langmead-lab.org/teaching-materials) to tell me briefly how you are using the slides. For original Keynote files, email me (ben.langmead@gmail.com).

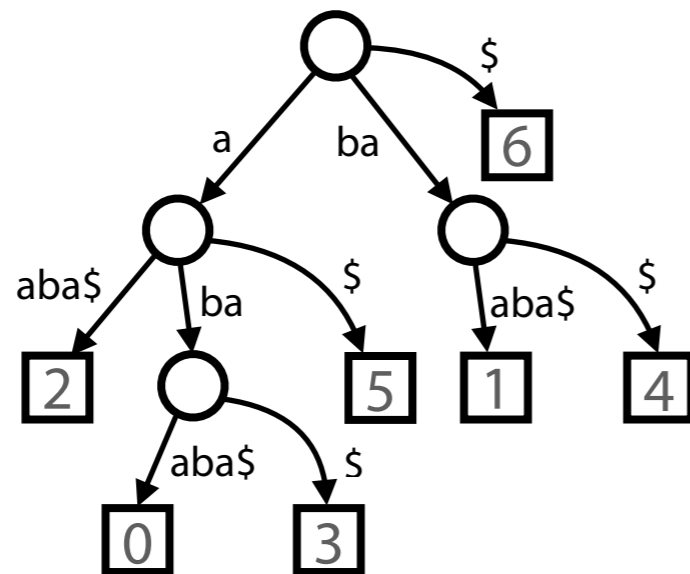
Indexing with suffixes

Still indexing with suffixes of T , but can we get smaller than the suffix tree?



Indexing with suffixes

Still indexing with suffixes of T , but can we get smaller than the suffix tree?



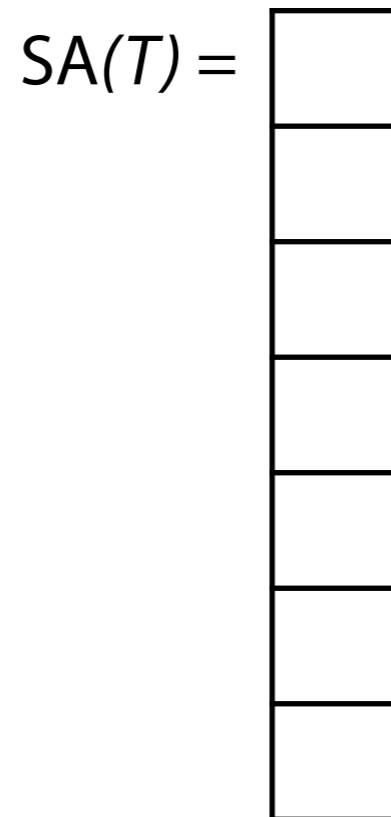
Suffix Tree

6	\$
5	a\$
3	aaba\$
1	aba\$
0	abaaba\$
4	ba\$
2	baaba\$

Suffix Array

Suffix array

$T = \text{abaaba\$}$
0123456



Array of integers $\in [0, m)$ in order according to lexicographic (alphabetical) order of T 's suffixes + T itself

Suffix array

$T = \text{abaaba\$}$
0123456

SA(T) =

6	\$
5	a \$
2	a a b a \$
3	a b a \$
0	a b a a b a \$
4	b a \$
1	b a a b a \$

m integers

Array of integers $\in [0, m)$ in order according to lexicographic (alphabetical) order of T 's suffixes + T itself

Suffix array

Space bound?

$T = \mathbf{abaaba\$}$

$SA(T) =$

6
5
2
3
0
4
1

Suffix array

Space bound?

$T = \mathbf{abaaba\$}$

m integers, m characters

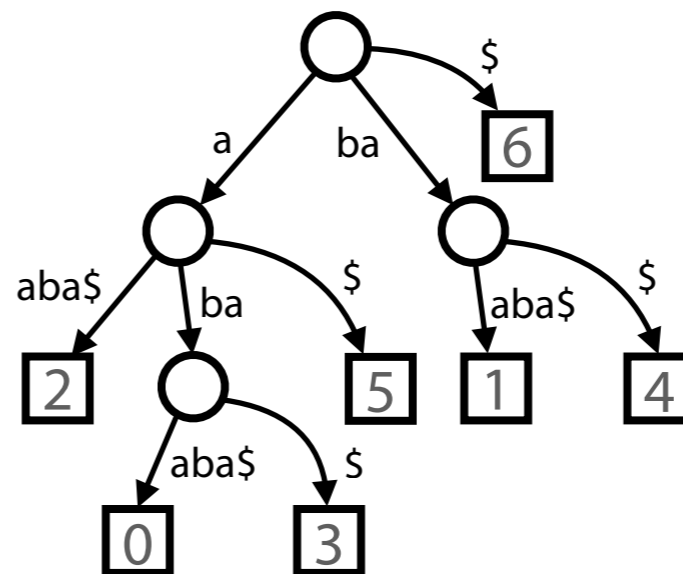
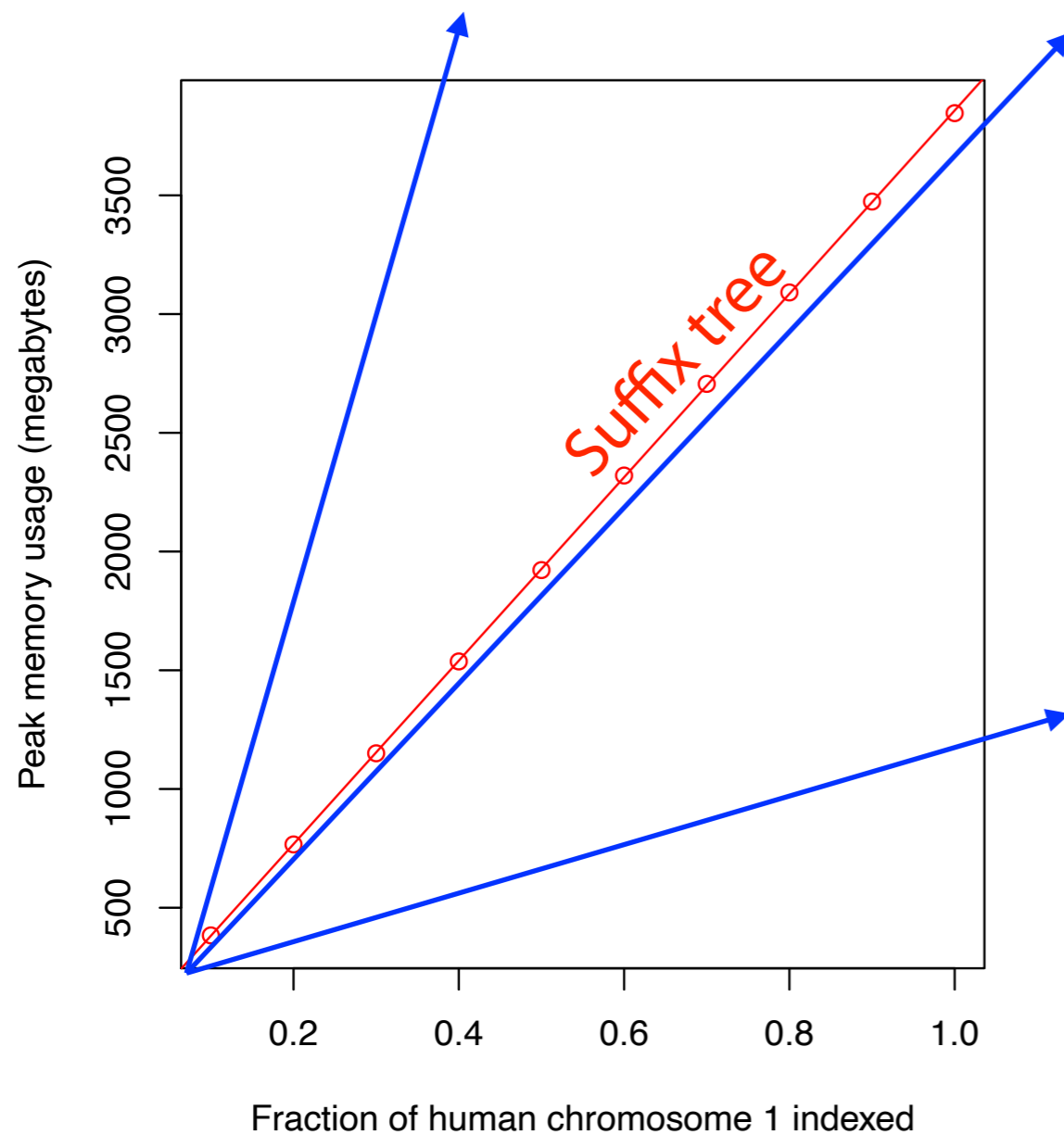
$SA(T) =$

6
5
2
3
0
4
1

$O(m)$ space, like suffix tree 

Suffix array

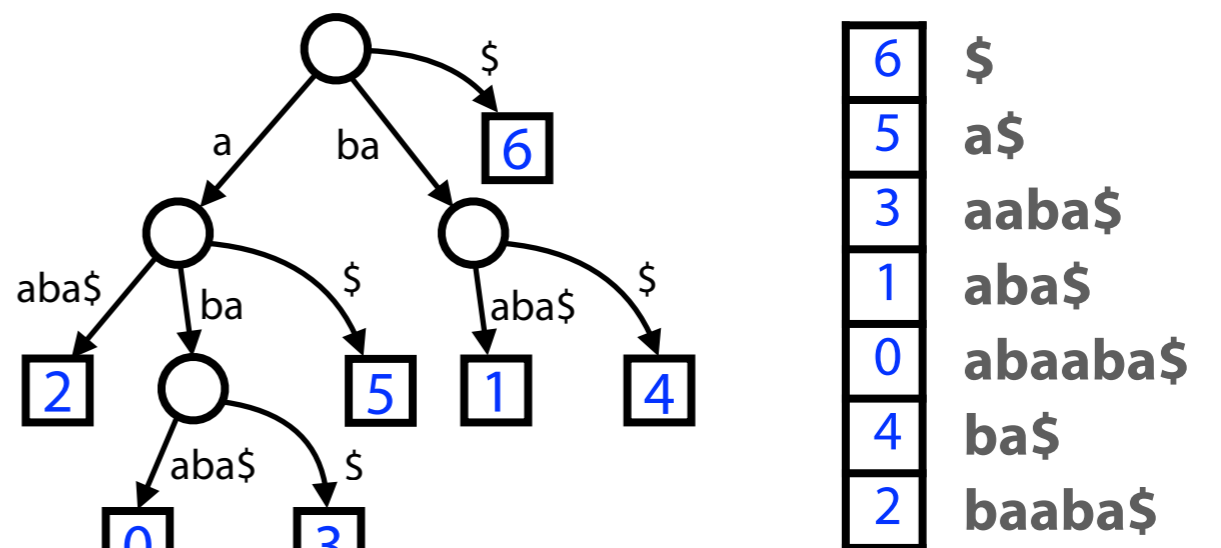
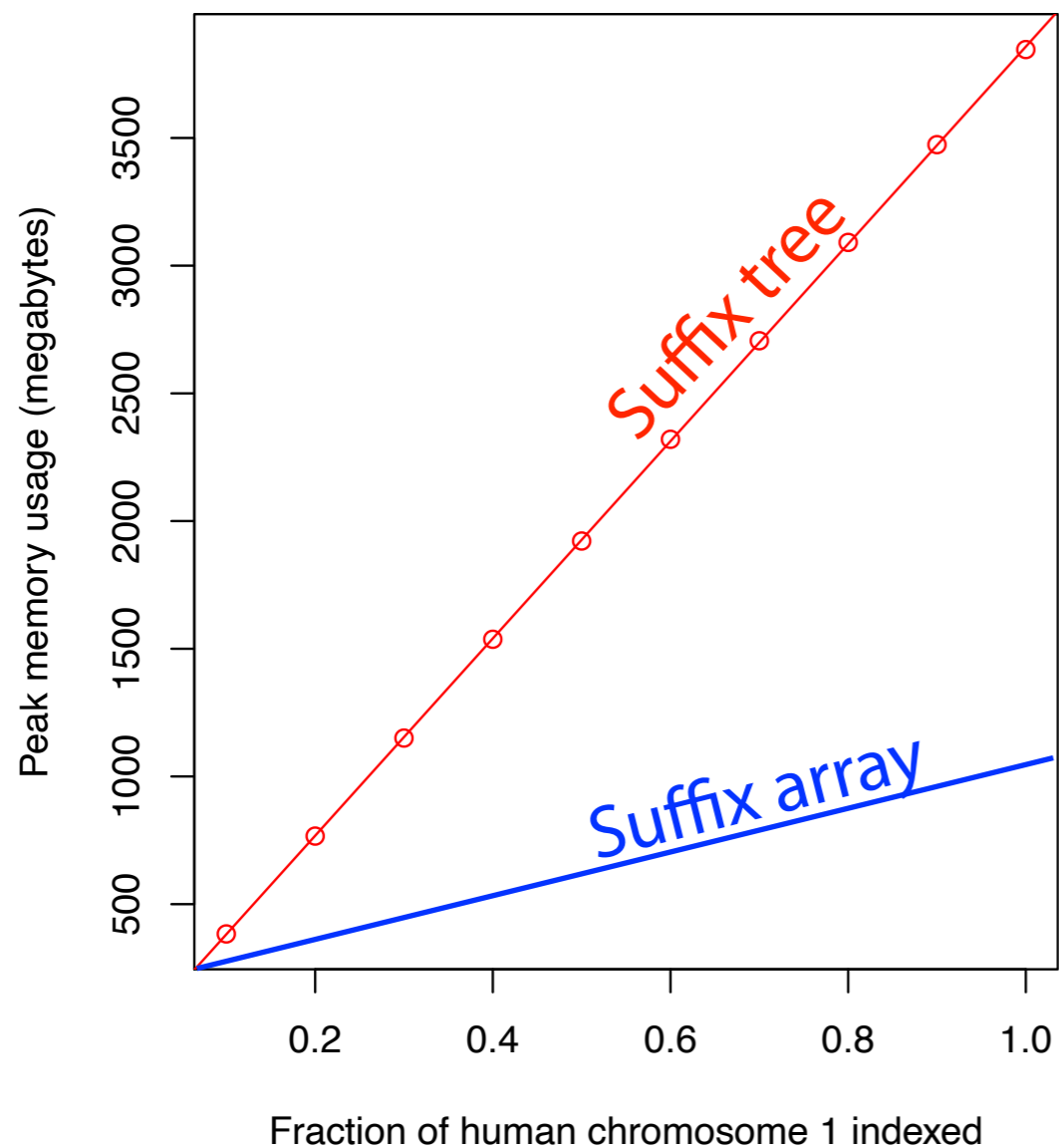
$O(m)$ space, but... Is "constant factor" worse, better, same?



6	\$
5	a\$
3	aaba\$
1	aba\$
0	abaaba\$
4	ba\$
2	baaba\$

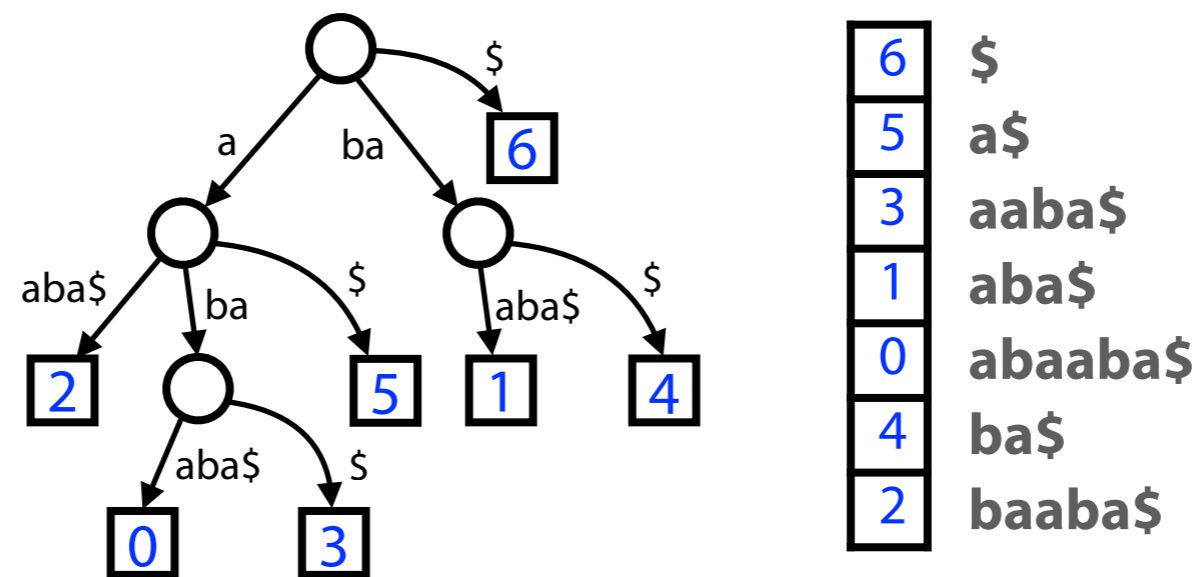
Suffix array

Leaves of suffix tree equal size of suffix array. Internal nodes, etc, are extra, making suffix tree bigger.



Suffix array

For human genome: suffix array consists of ~3 billion 32-bit integers \approx 12 GB. (Plus T)



Suffix tree will be >45 GB, possibly much larger depending on implementation

Kurtz, Stefan. "Reducing the space requirement of suffix trees."
Software: Practice and Experience 29.13 (1999): 1149-1171.