Suffix Arrays
Ben Langmead

Please sign guestbook ([www.langmead-lab.org/teaching-materials](http://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](mailto:ben.langmead@gmail.com)).
Indexing with suffixes

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

**Suffix trie**

>100K nodes

**Suffix tree**

<1K nodes
Indexing with suffixes

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

Suffix Tree

Suffix Array
Suffix array

\[ T = \text{abaaba$} \]
\[ \text{SA}(T) = \]

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

\[ T = \text{abaaba}\$

\[ 0123456 \]

\[ SA(T) = \]

\[ \begin{array}{c|c}
0 & a b a a b a \$ \\
1 & b a a b a \$ \\
2 & a a b a \$ \\
3 & a b a \$ \\
4 & b a \$ \\
5 & a \$ \\
6 & \$ \\
\end{array} \]

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

Space bound?

$T = \text{abaaba}\$

$$SA(T) = \begin{array}{c}
6 \\
5 \\
2 \\
3 \\
0 \\
4 \\
1 
\end{array}$$
Suffix array

Space bound?

\[ m \text{ integers, } m \text{ characters} \]

\[ T = \text{abaaba}\$

\[ \text{SA}(T) = \begin{array}{c} 6 \\ 5 \\ 2 \\ 3 \\ 0 \\ 4 \\ 1 \end{array} \]

\[ O(m) \text{ space, like suffix tree} \]
Suffix array

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