Suffix Tries: size

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).

Suffix trie

Build a **trie** containing all **suffixes** of a text *T*

```
T: GTTATAGCTGATCGCGGCGTAGCGG$
   GTTATAGCTGATCGCGGCGTAGCGG$
    TTATAGCTGATCGCGGCGTAGCGG$
     TATAGCTGATCGCGGCGTAGCGG$
      ATAGCTGATCGCGGCGTAGCGG$
       TAGCTGATCGCGGCGTAGCGG$
        AGCTGATCGCGGCGTAGCGG$
         GCTGATCGCGGCGTAGCGG$
          CTGATCGCGGCGTAGCGG$
           TGATCGCGGCGTAGCGG$
            GATCGCGGCGTAGCGG$
             ATCGCGGCGTAGCGG$
              TCGCGGCGTAGCGG$
               CGCGGCGTAGCGG$
                GCGGCGTAGCGG$
                 CGGCGTAGCGG$
                  GGCGTAGCGG$
                   GCGTAGCGG$
                    CGTAGCGG$
                     GTAGCGG$
                      TAGCGG$
                       AGCGG$
                        GCGG$
                         CGG$
                          GG$
                           G$
                            $
```

m(m+1)/2 chars

Suffix trie

How does the suffix trie grow with $| T | = m$?

T: **a b a a b a $**

b a a b a $

a a b a $

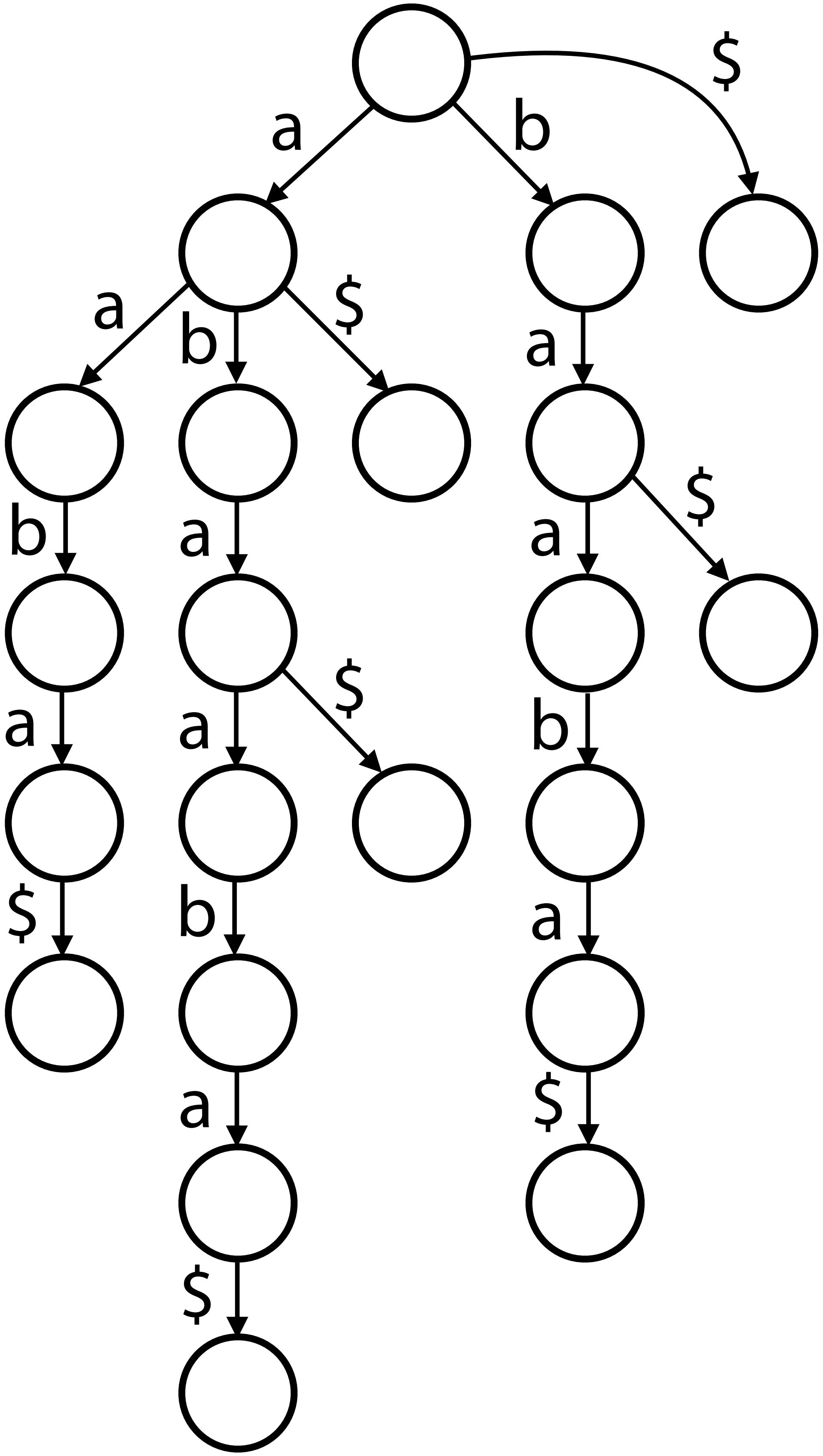
a b a $

b a $

a $

$

Does "prefix sharing" save us?



Suffix trie

Take repetitive strings of the form T = aaaaaa$ (a^m$)

T: **a a a a a a $**

a a a a a $

a a a a $

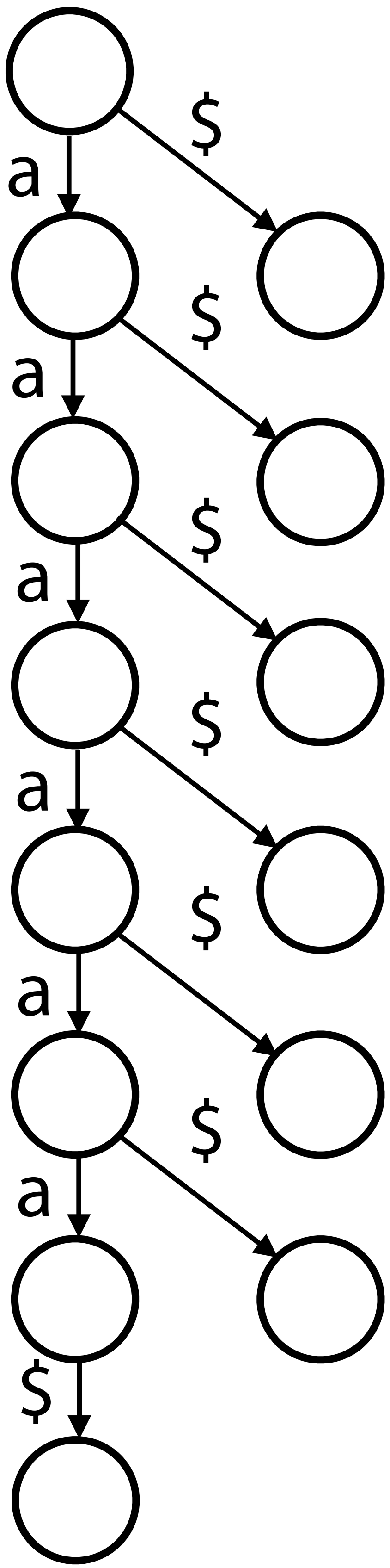
a a a $

a a $

a $

$

Growth is O(*m*), thanks to prefix sharing

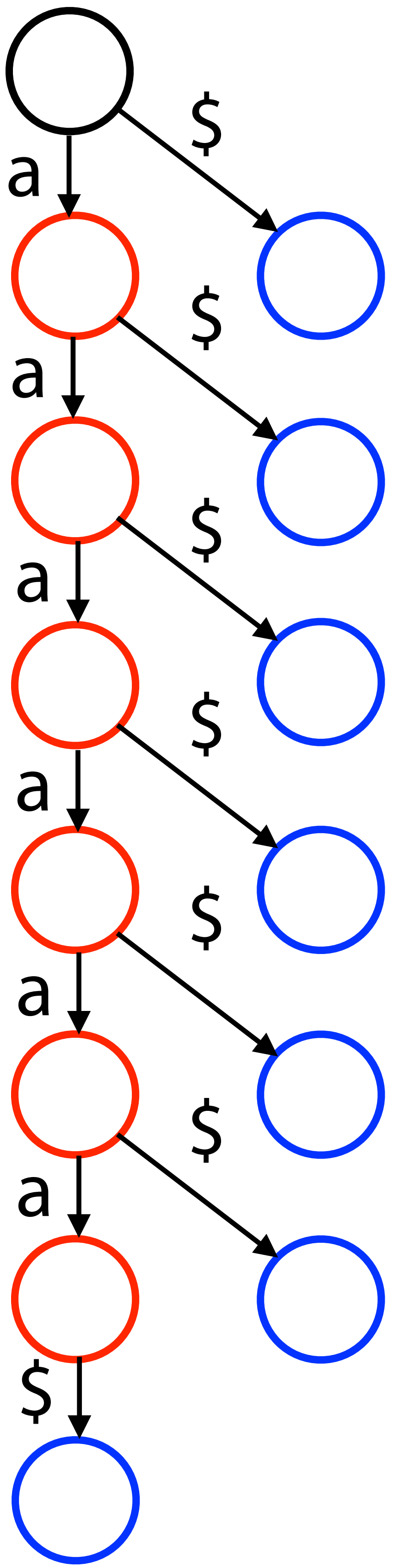


Suffix trie

Take repetitive strings of the form T = aaaaaa$ (a^m\$)

T: aaaaaa$
aaaaa$
aaaa$
aaa$
aa$
a$
$

Growth is O(*m*), thanks to prefix sharing



Suffix trie

Can suffixes have **no** prefix sharing?

Yes: all distinct characters

T: **a b c d e f $**

b c d e f $

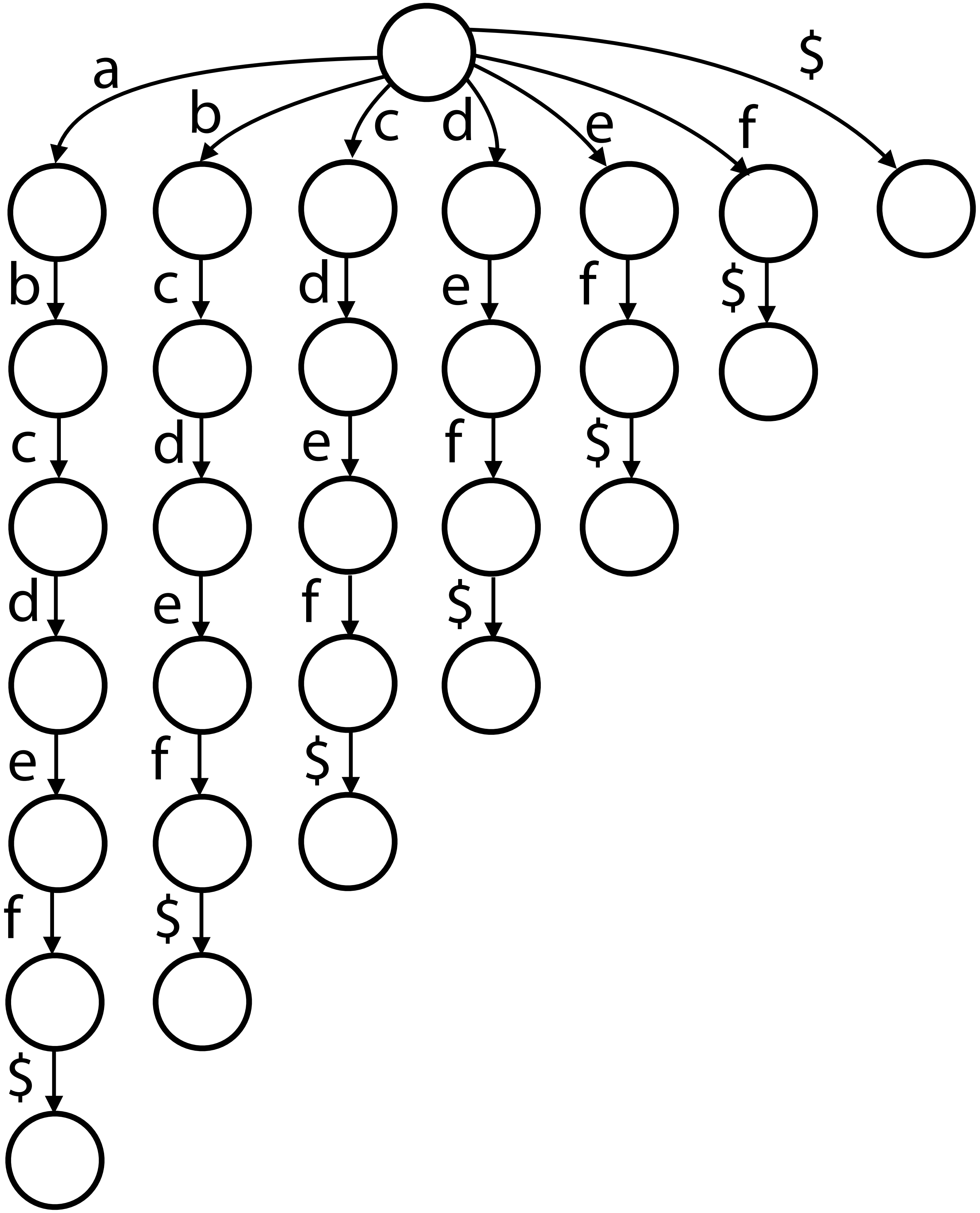
c d e f $

d e f $

e f $

f $

$



Suffix trie

Can suffixes have **no** prefix sharing?

Yes: all distinct characters

T: **a b c d e f $**

b c d e f $

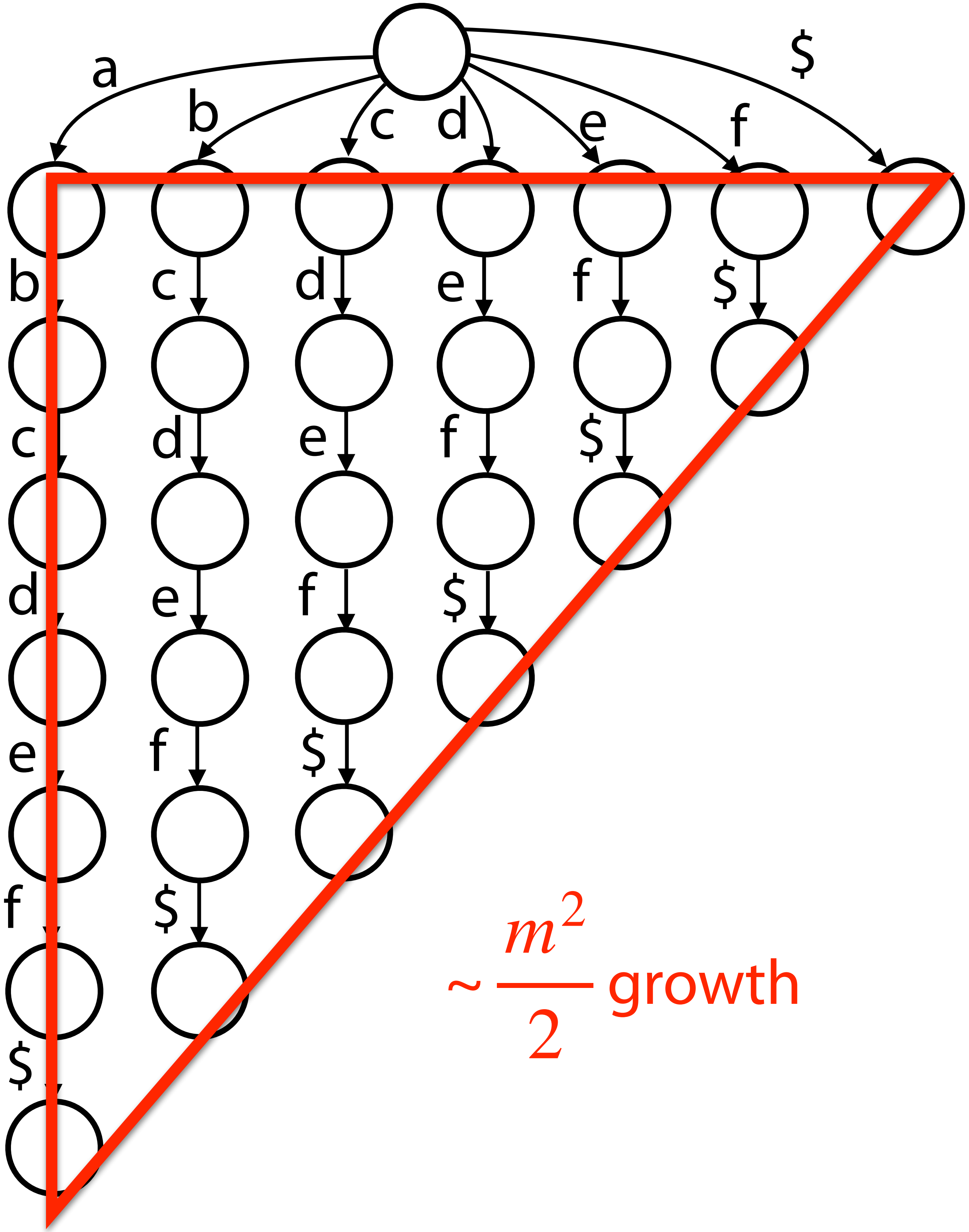
c d e f $

d e f $

e f $

f $

$



$\sim \frac{m^2}{2}$ growth

Suffix trie

Even when alphabet is {a, b}, we can find strings where suffix trie grows with O(m^2)

Suffix trie

Even when alphabet is {a, b},
we can find strings where
suffix trie grows with O(m^2)

T: **a a a b b b $**

Suffix trie

Even when alphabet is {a, b},
we can find strings where
suffix trie grows with O(m^2)

T: **a a a b b b $**

a a b b b $

a b b b $

b b b $

b b $

b $

$

Suffix trie

Even when alphabet is {a, b}, we can find strings where suffix trie grows with $O(m^2)$

T: **a a a b b b $**

a a b b b $

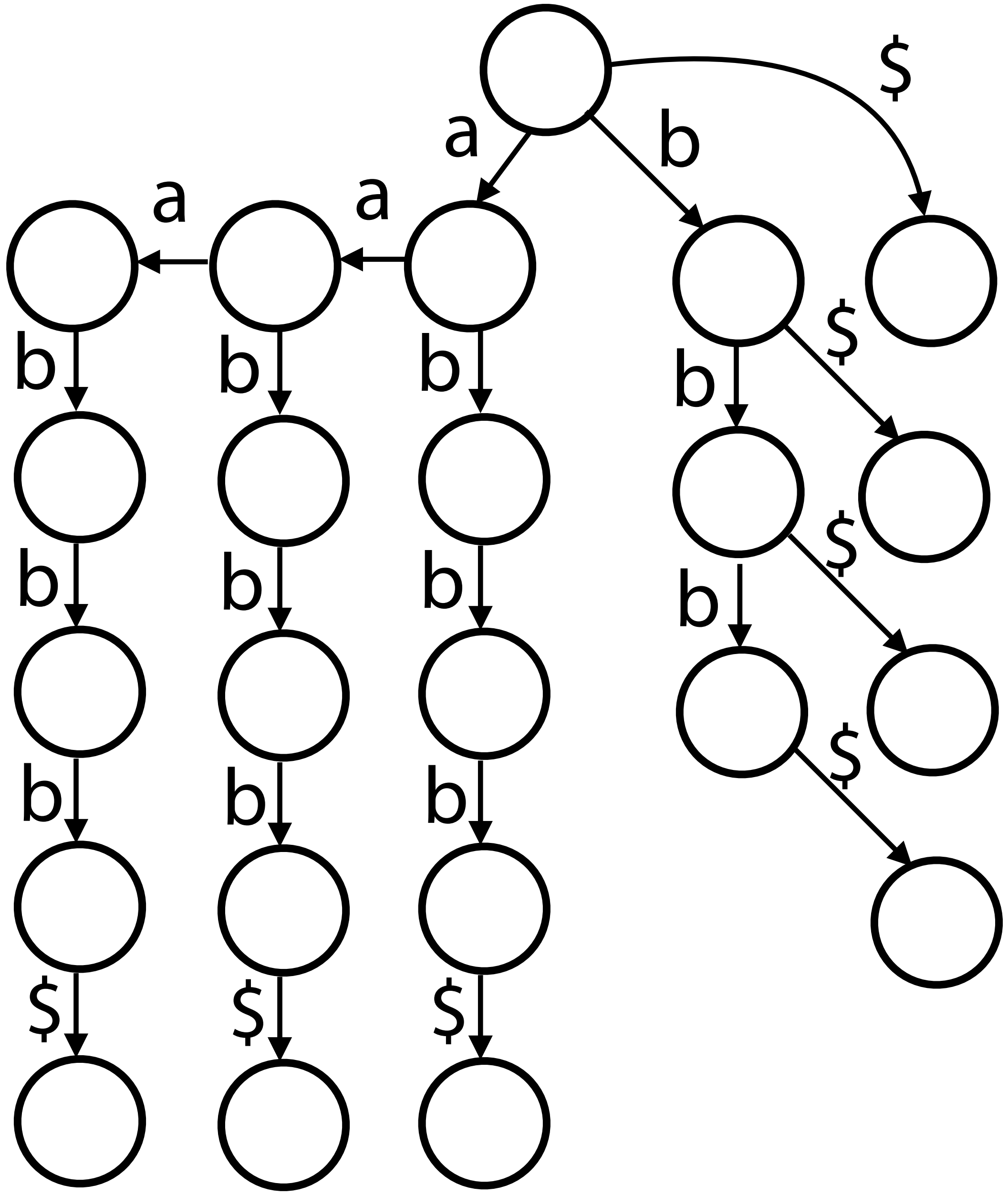
a b b b $

b b b $

b b $

b $

$



Suffix trie

Even when alphabet is {a, b}, we can find strings where suffix trie grows with O(m²)

T: **a a a b b b $**

a a b b b $

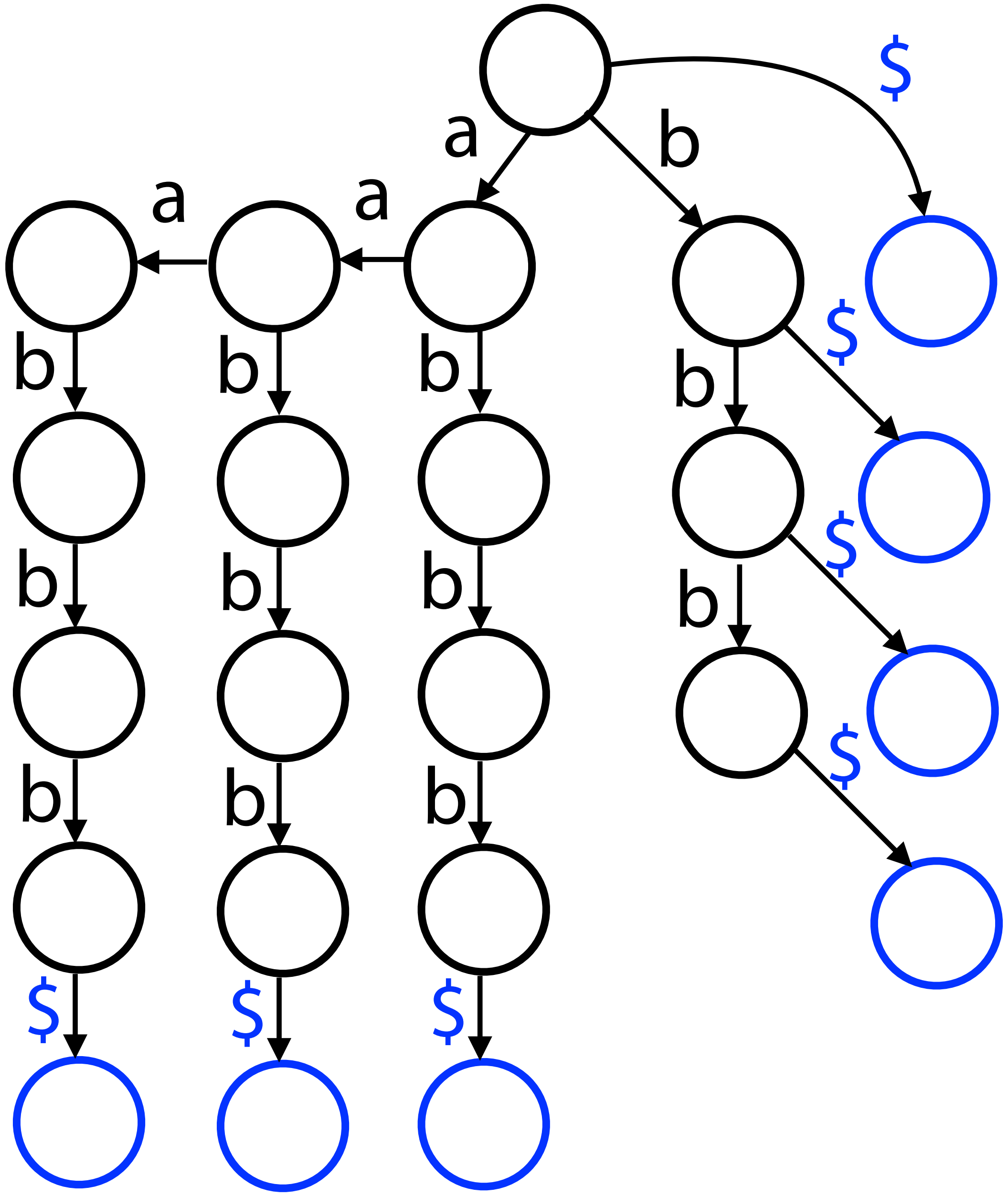
a b b b $

b b b $

b b $

b $

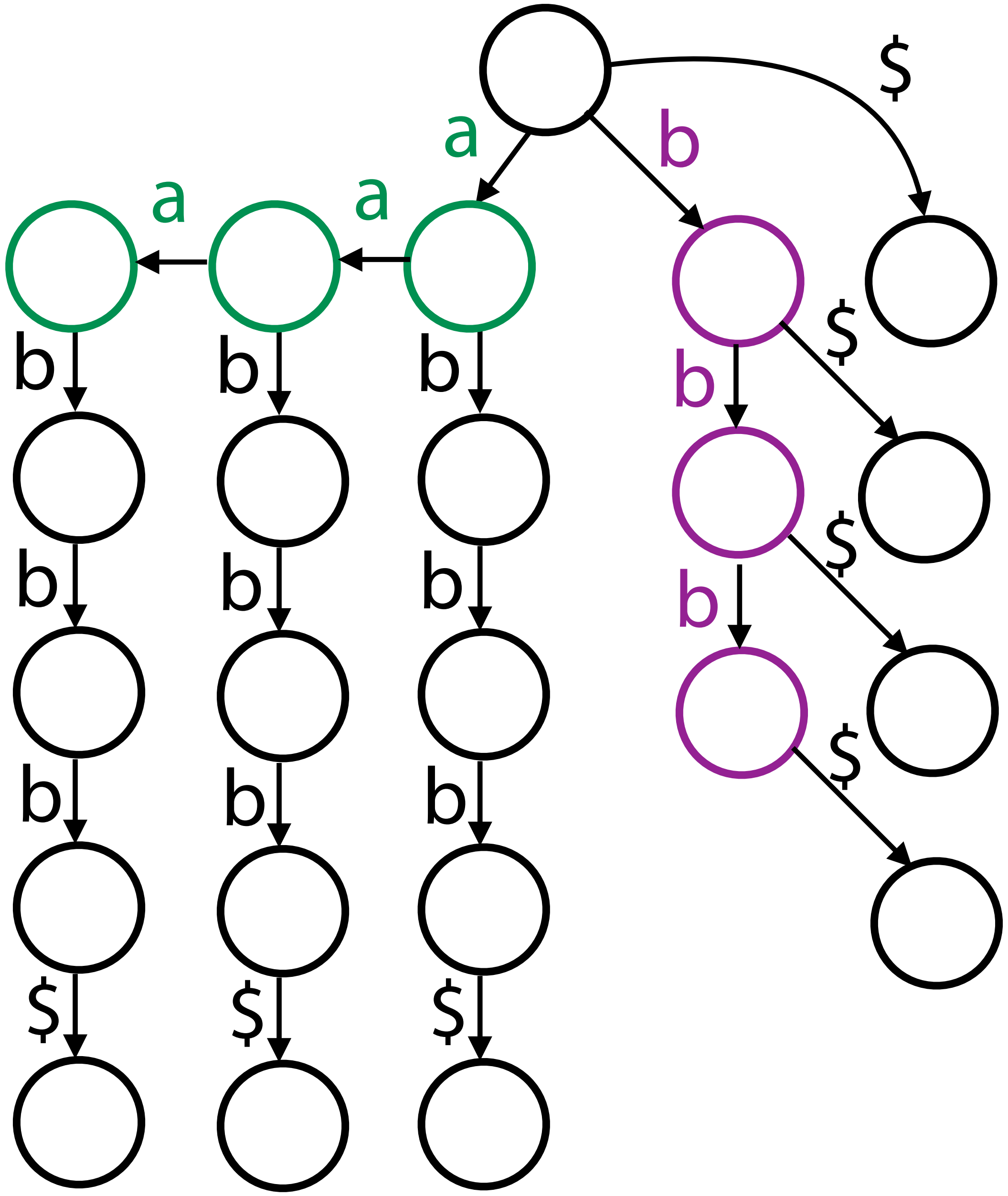
$



Suffix trie

Even when alphabet is {a, b}, we can find strings where suffix trie grows with O(m^2)

T: **a a a b b b $**
a a b b b $
a b b b $
b b b $
b b $
b $
$



Suffix trie

Even when alphabet is {a, b}, we can find strings where suffix trie grows with O(m^2)

T: a a a b b b $

a a b b b $

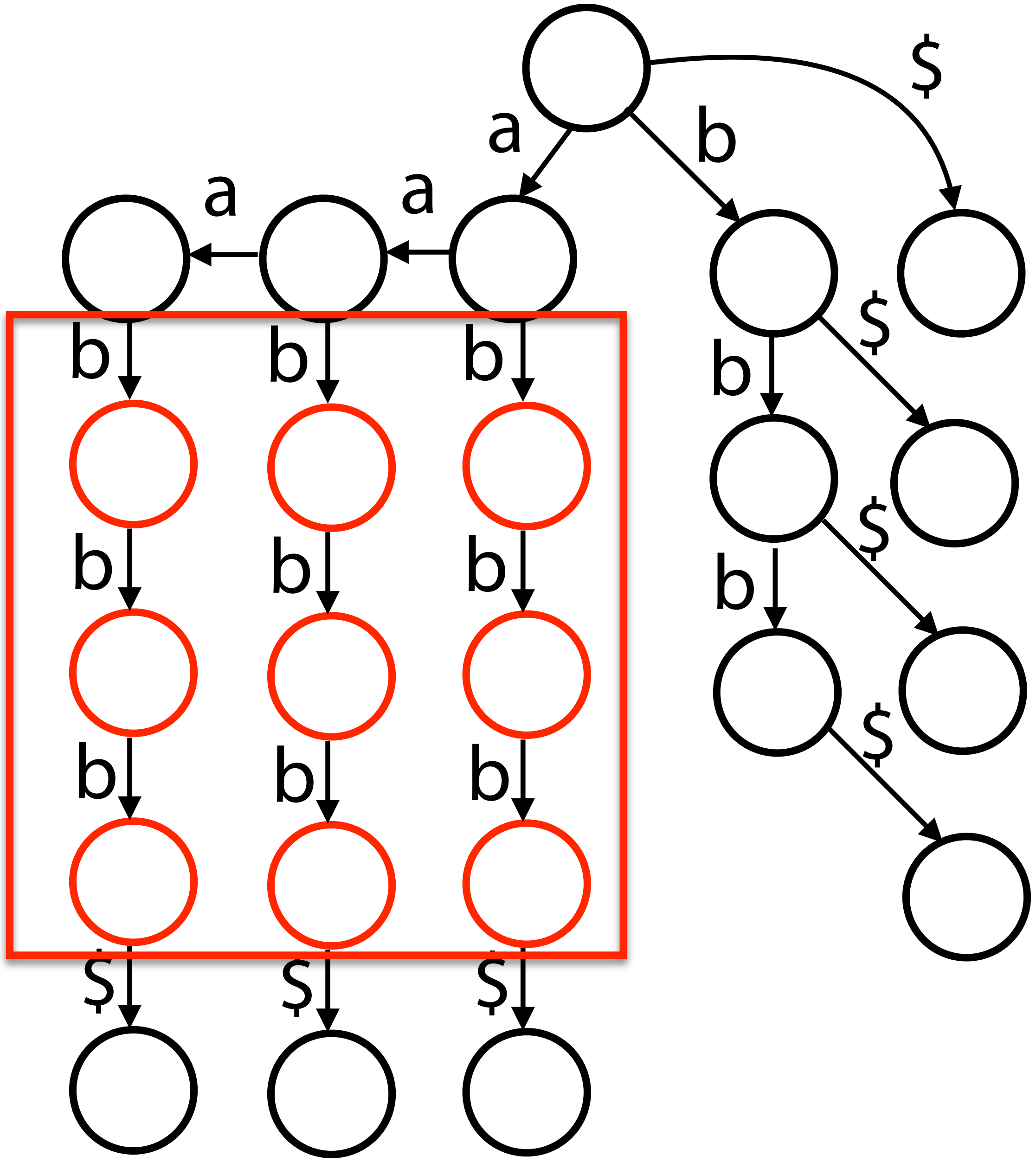
a b b b $

b b b $

b b $

b $

$

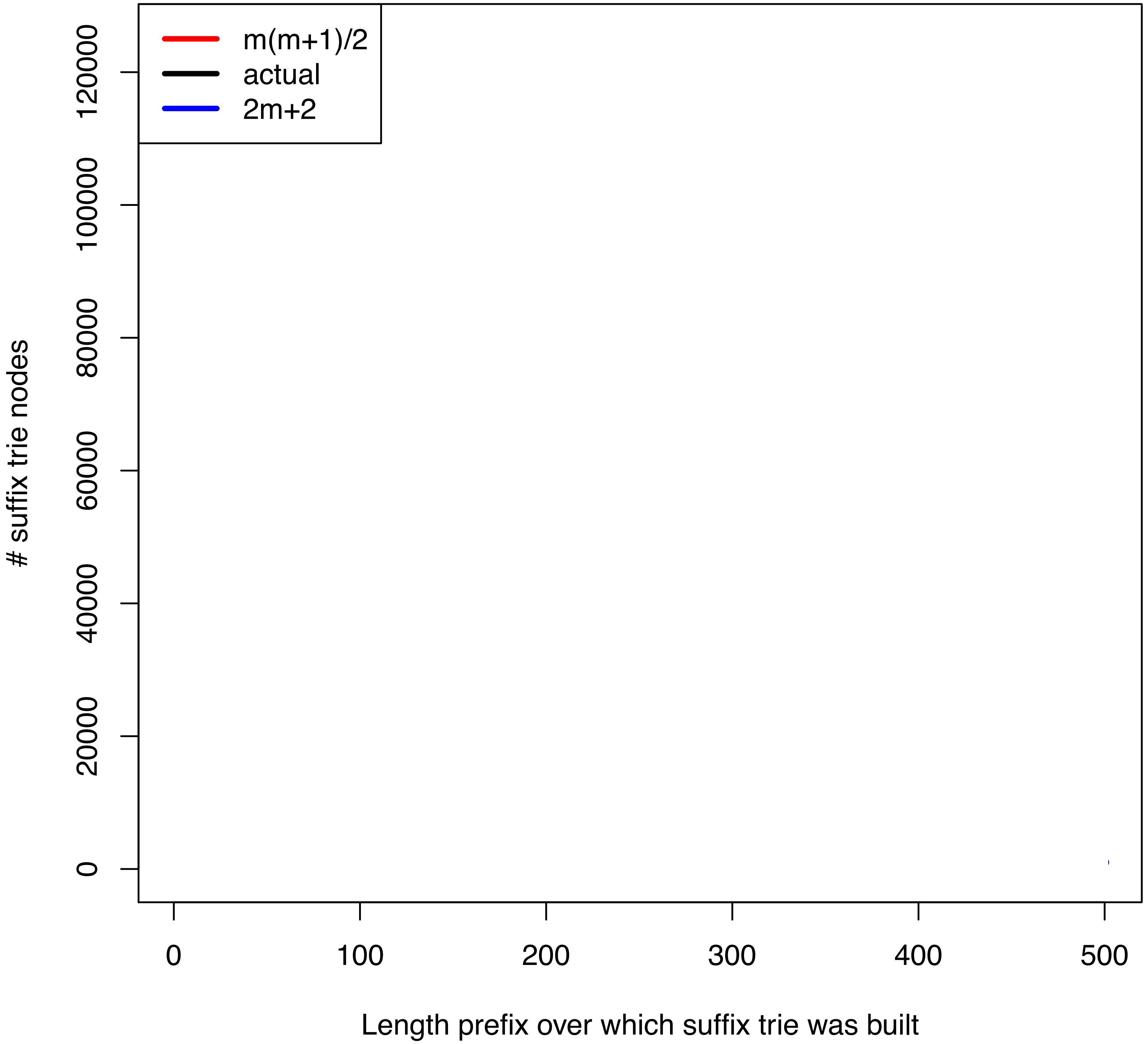


$$\sim\left(\frac{m}{2}\right)^2$$

Suffix trie: actual growth

Built suffix tries for the first 500 prefixes of the lambda phage virus genome

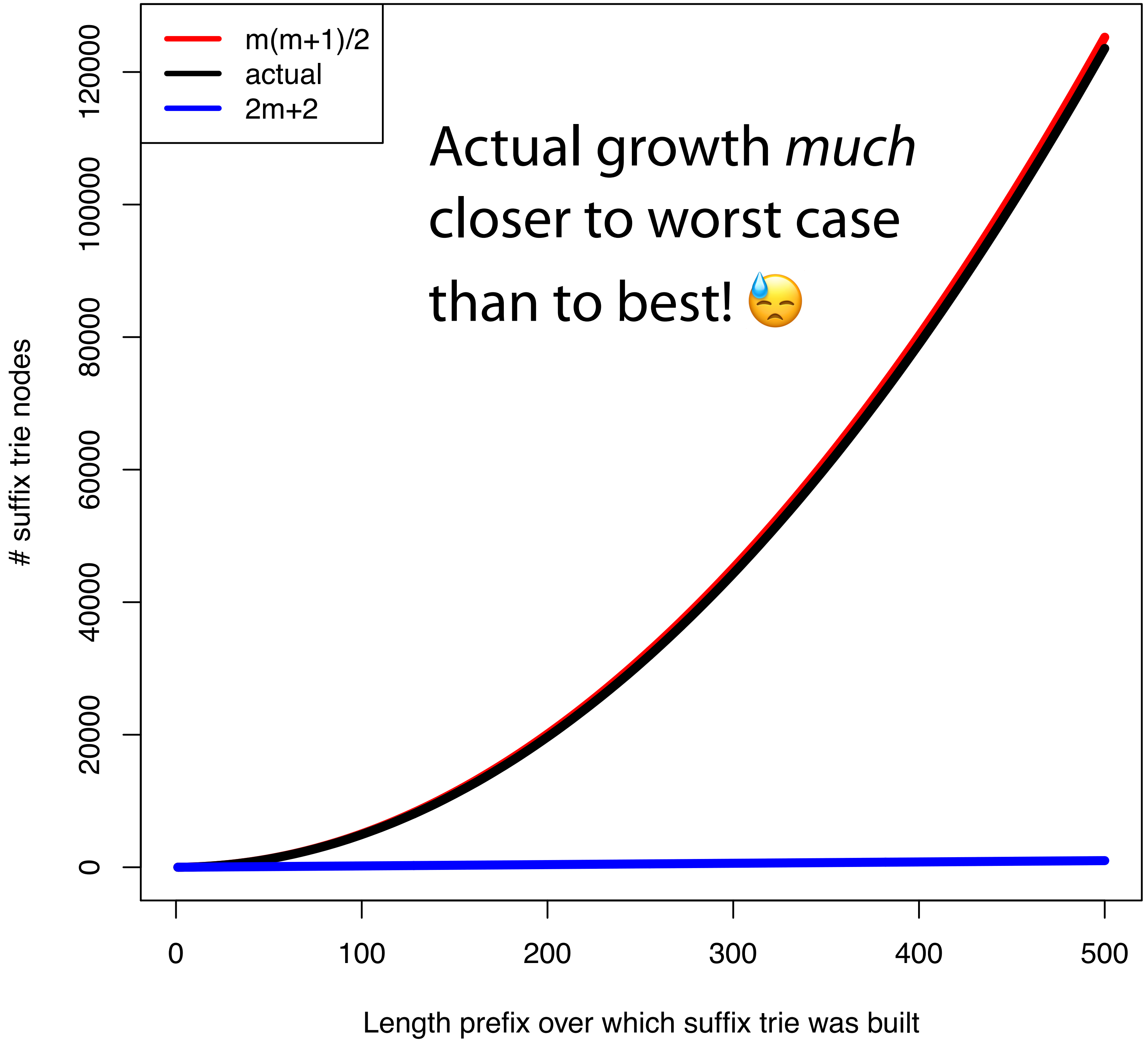
Black curve shows how # nodes increases with prefix length



Suffix trie: actual growth

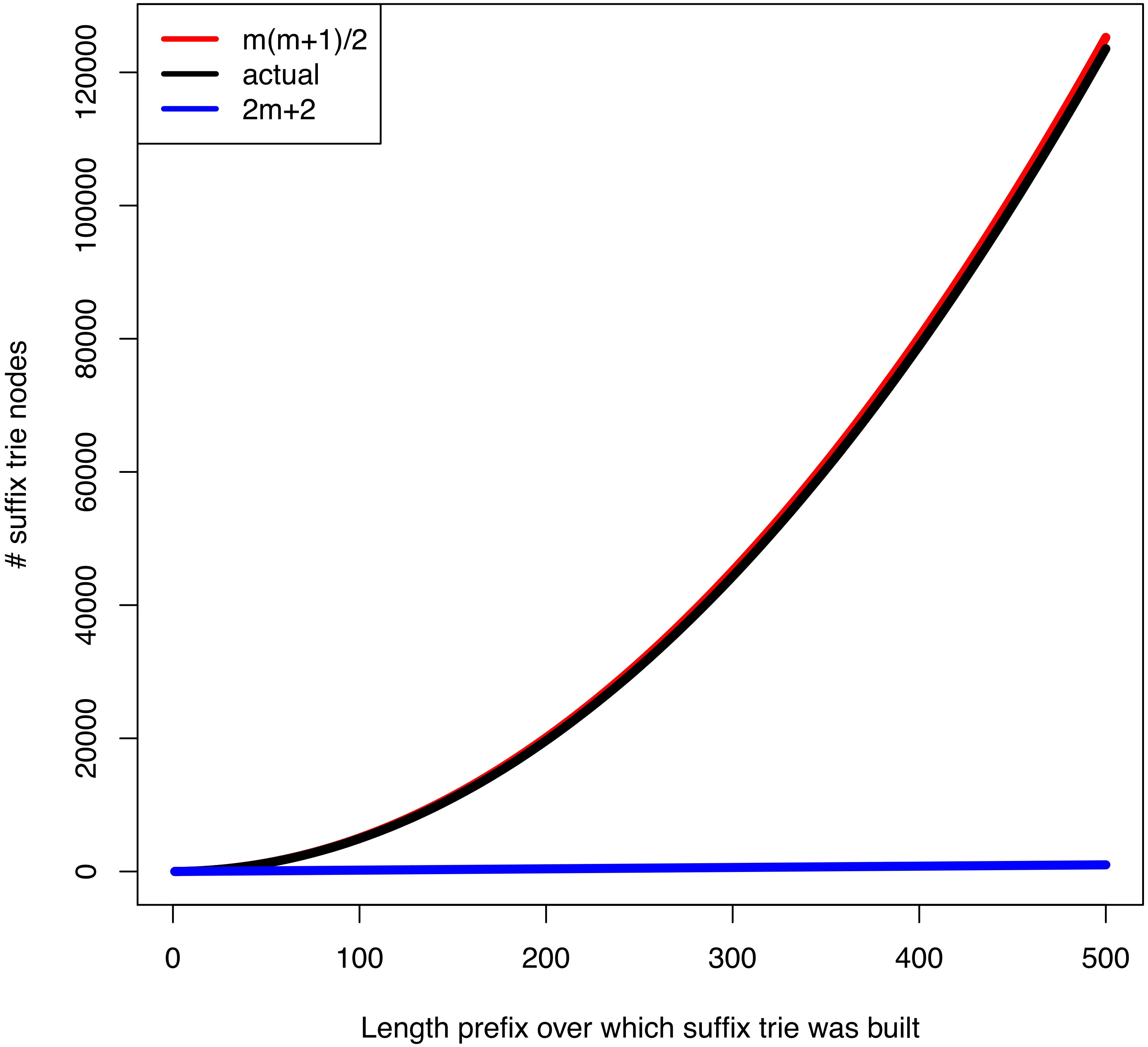
Built suffix tries for the first 500 prefixes of the lambda phage virus genome

Black curve shows how # nodes increases with prefix length



Actual growth *much* closer to worst case than to best! 😓

Suffix trie: actual growth



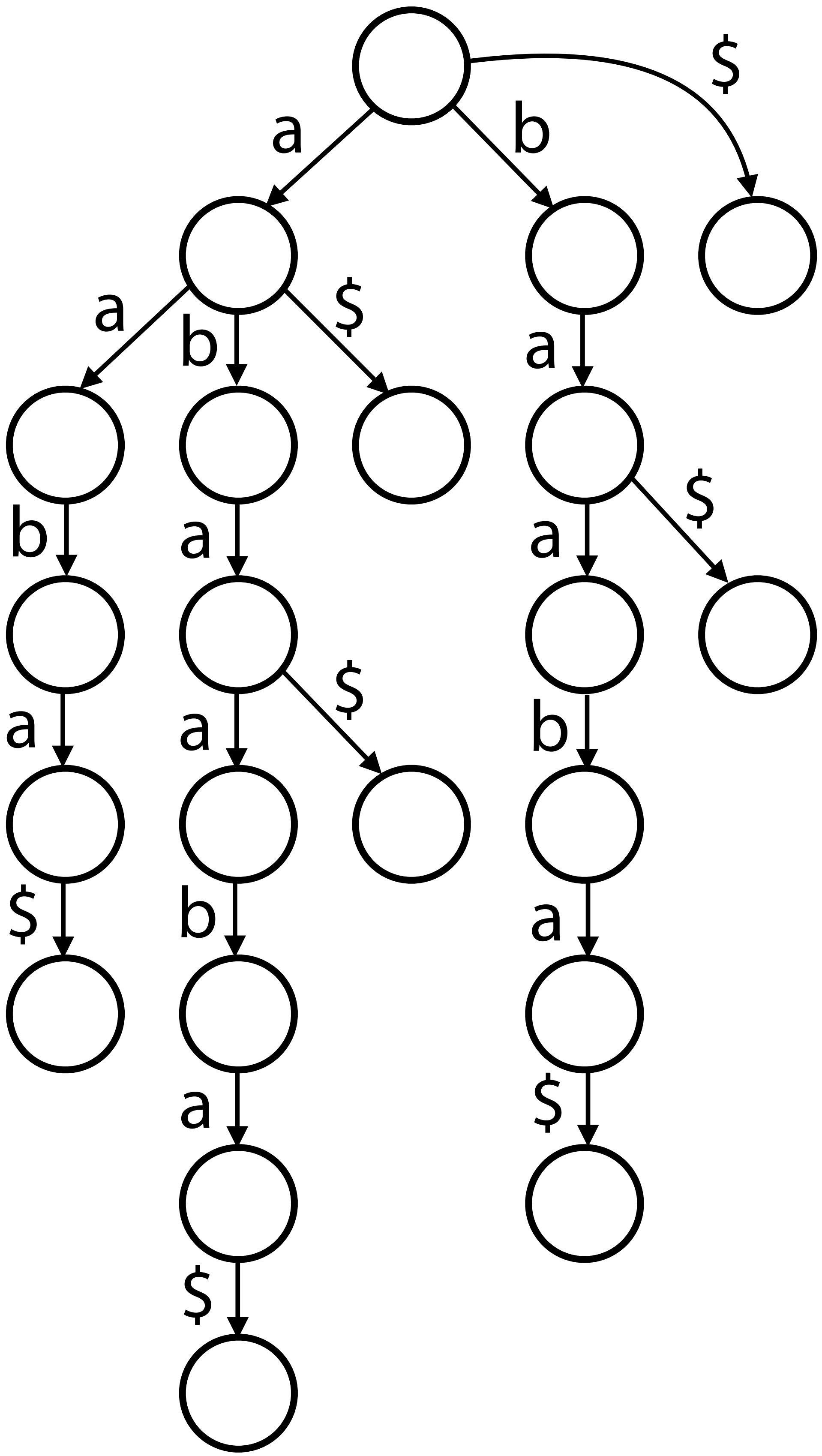
Human genome is $3 \cdot 10^9$ bases long

If $m = 3 \cdot 10^9$, m^2 is far beyond what we can store in memory

Suffix trie

How do we **shrink** the trie?

In next video...


a
b
$
a
b
$
a
$
b
a
a
a
$
a
a
$
$
b
b
a
a
$
$