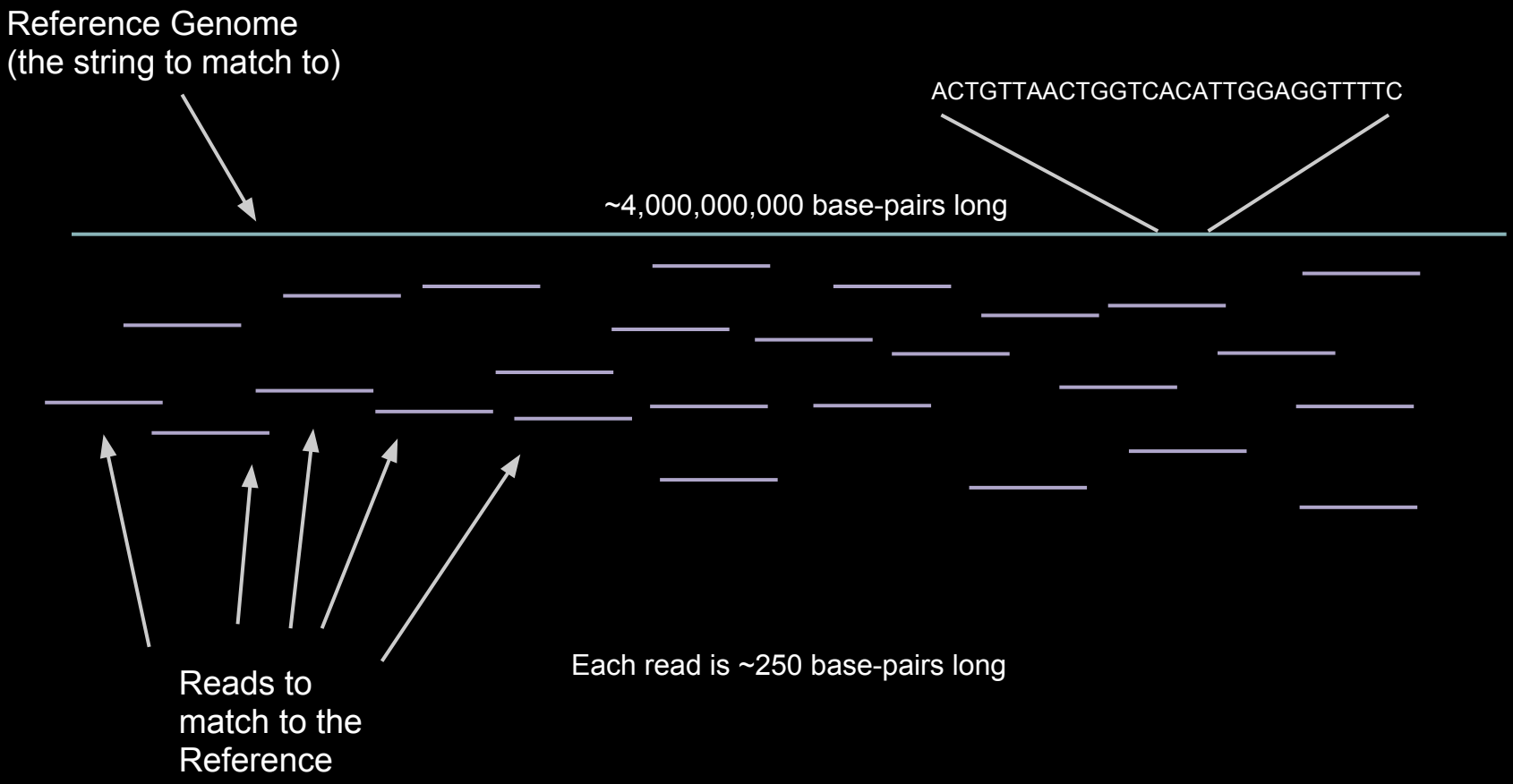
Suffix Structures in Computational Biology

Presented by: Cole Lyman

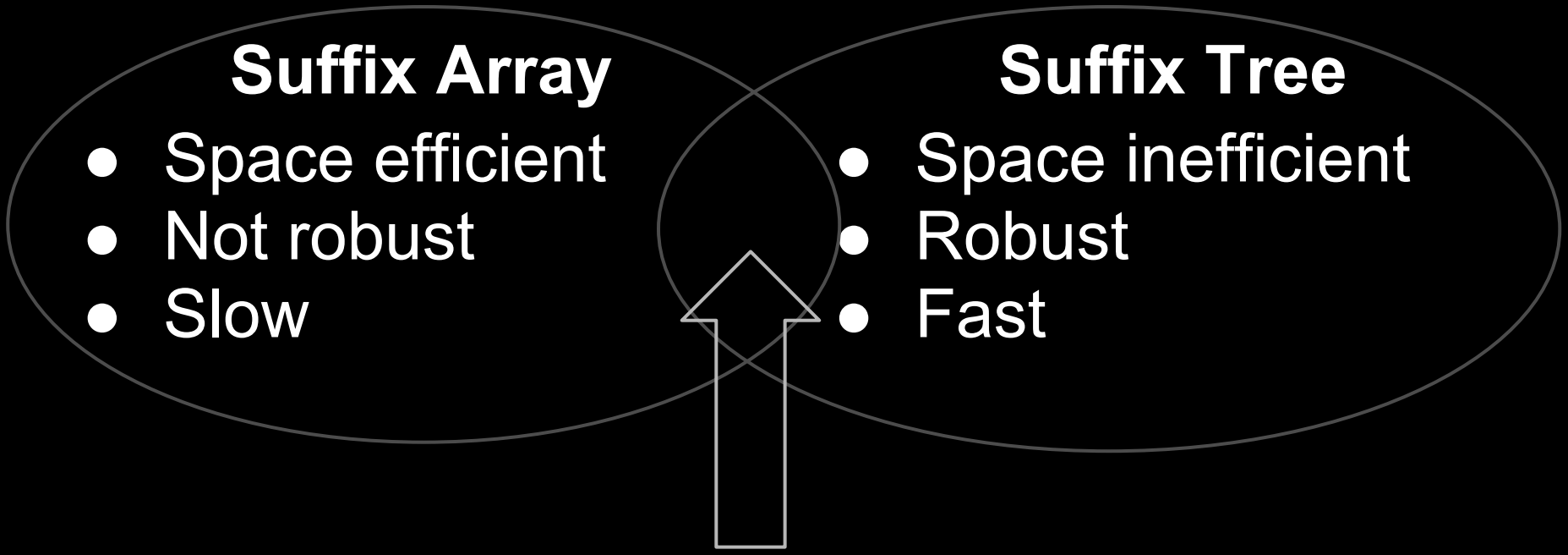
Suffix Structures

- Data structures where all possible suffixes of a string are represented.
- *O(n):*
  - Construction
  - Searching for suffixes
- Used in string matching
  - Genome mapping

Genome Mapping



Types of Suffix Structures



Suffix Cactus

Space efficient yet robust

Ideal choice for Computational Biology

Example of Suffix Array

Suffix Array, *A,* of string, ***banana$.***

0 1 2 3 4 5 6 7

Index of string	Suffix
0	banana$
1	anana$
2	nana$
3	ana$
4	na$
5	a$
6	$

Sort

Index of string	Suffix
6	$
5	a$
3	ana$
1	anana$
0	banana$
4	na$
2	nana$

space optimization

equivalent

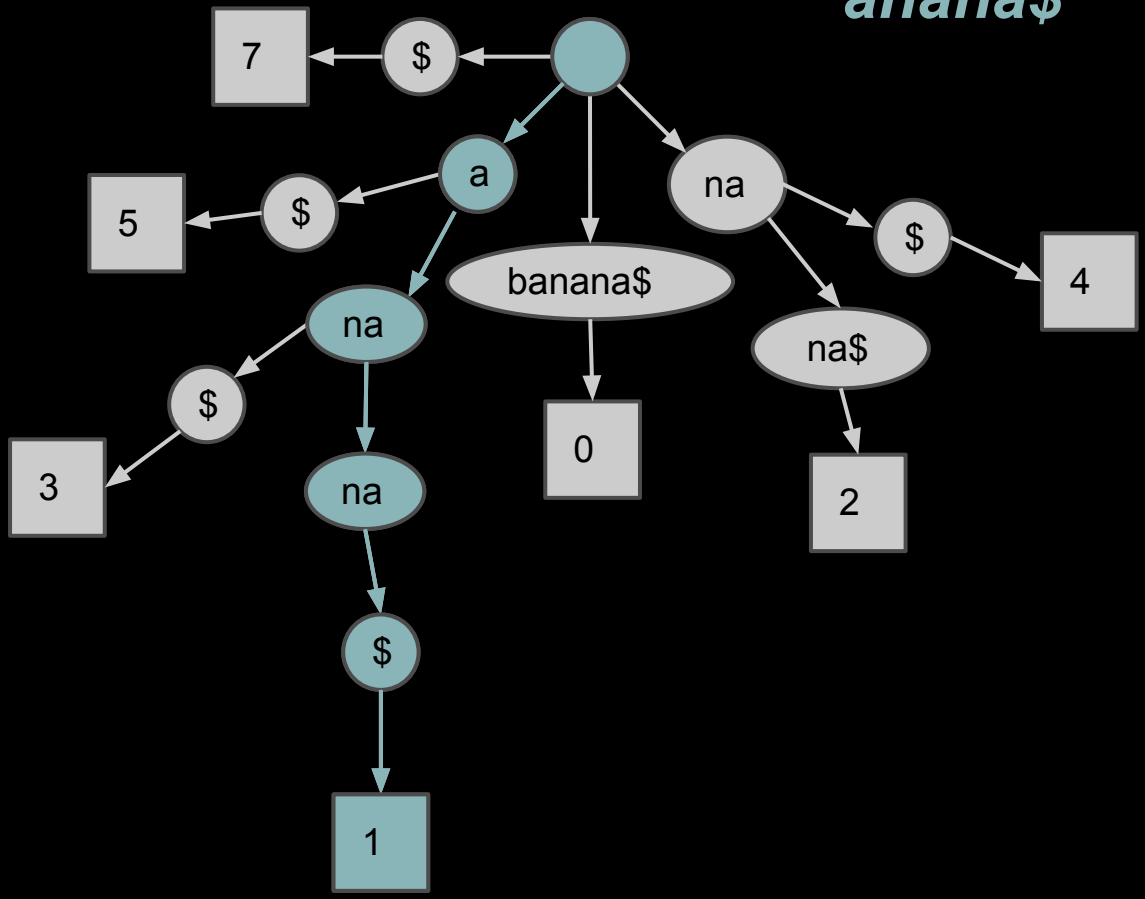
Position of array	Index of string
0	6
1	5
2	3
3	1
4	0
5	4
6	2

Example of Suffix Tree

0 1 2 3 4 5 6 7

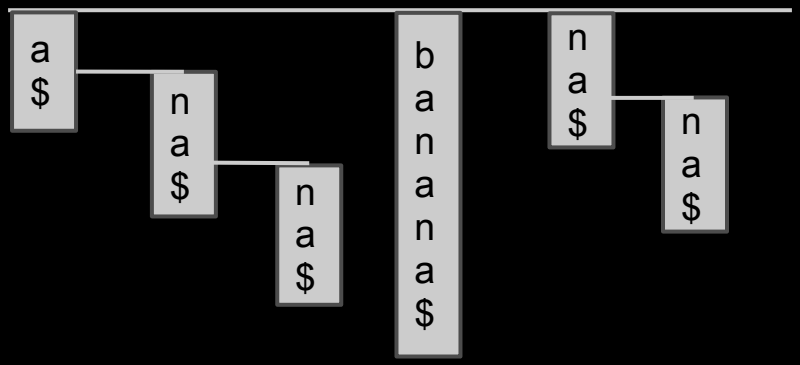
Suffix Tree, *T,* of string, ***banana$.***

anana$



Example of Suffix Cactus

Suffix Cactus, *C,* of string, ***banana$.***



Comparison of Structures

Measurements are based off of DNA sequences with 300,000 bases.

Space and Construction Time*

Type of Structure	Space (bytes/*n*)	Time (*s*)
Tree	17.70	5.62
Array	10.00	41.40
Cactus via Tree	18.70	7.78
Cactus via Array	10.00	42.60

String Matching Time*

Type of Structure	Time (*s*)
Tree	0.96
Array	0.71
Cactus	0.61

*Data from: Kärkkäinen, Juha. *Suffix Cactus: A Cross between Suffix Tree and Suffix Array*. Page 197. Springer. 1995.

Comparison of Structures

Measurements are based off of an english sequence with 300,000 bases, and 77 characters in the alphabet.

Space and Construction Data*

Type of Structure	Space (bytes/*n*)	Time (*s*)
Tree	15.17	6.60
Array	10.00	36.40
Cactus via Tree	16.17	8.63
Cactus via Array	10.00	37.7

String Matching Data*

Type of Structure	Time (*s*)
Tree	1.63
Array	0.67
Cactus	1.86

*Data from: Kärkkäinen, Juha. *Suffix Cactus: A Cross between Suffix Tree and Suffix Array*. Page 197. Springer. 1995.

Comparison of Structures

Measurements are based off of a random sequence with 300,000 bases and 16 characters in the alphabet.

Space and Construction Time*

Type of Structure	Space (bytes/*n*)	Time (*s*)
Tree	11.80	8.10
Array	10.00	31.20
Cactus via Tree	12.80	9.91
Cactus via Array	10.00	32.50

String Matching Time*

Type of Structure	Time (*s*)
Tree	0.66
Array	0.63
Cactus	0.90

*Data from: Kärkkäinen, Juha. *Suffix Cactus: A Cross between Suffix Tree and Suffix Array*. Page 197. Springer. 1995.