3. Gene prediction

- All genes end on a stop codon
- A simple algorithm for gene prediction
- Searching for start and stop codons
- Predicting all the genes in a sequence
- Making the predictions more reliable
- Boyer-Moore algorithm
- **Index and suffix trees**
- Probabilistic methods
- Benchmarking the prediction methods
- Gene prediction in eukaryotic genomes

Improving the performance of string searching

- Preprocess the pattern
  - Boyer-Moore algorithm
- Preprocess the searchable text
  - Index of fixed-length words
  - Prefix tree

Index of fixed-length word

```
ACGGCTAGTTAGAA*
```

Index of fixed-length words

ACGGCTAGTTAGAA*

length = 14

Index of fixed-length words

ACGGCTAGTTAGAA*

length = 14

AA*	13
ACG	1
AGA	11
AGT	7
CGG	2
CTA	5
GAA	12
GCT	4
GGC	3
GTT	8
TAG	6, 10
TTA	9

Index of fixed-length words

ACGGCTAGTTAGAA*

TAG

AA*	13
ACG	1
AGA	11
AGT	7
CGG	2
CTA	5
GAA	12
GCT	4
GGC	3
GTT	8
TAG	6, 10
TTA	9

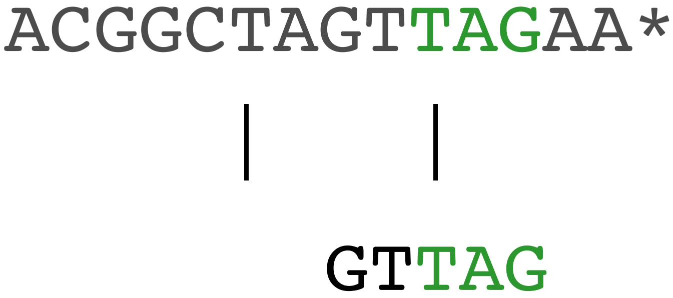
Index of fixed-length words

ACGGCTAGTTAGAA*

TAG

AA*	13
ACG	1
AGA	11
AGT	7
CGG	2
CTA	5
GAA	12
GCT	4
GGC	3
GTT	8
TAG	6, 10
TTA	9

Index of fixed-length words



ACGGCTAGTTAGAA*

GTTAG

TAG

AA*	13
ACG	1
AGA	11
AGT	7
CGG	2
CTA	5
GAA	12
GCT	4
GGC	3
GTT	8
TAG	6, 10
TTA	9

Index of fixed-length words

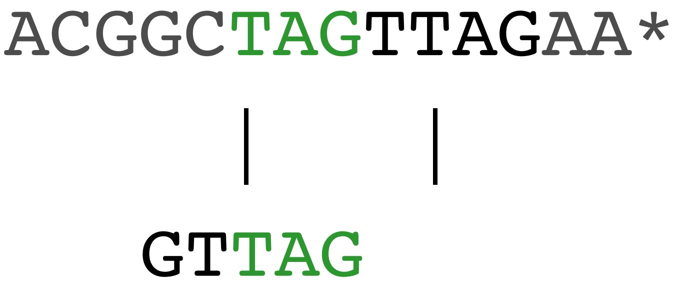
ACGGCTAGTTAGAA*

GTTAG

TAG

AA*	13
ACG	1
AGA	11
AGT	7
CGG	2
CTA	5
GAA	12
GCT	4
GGC	3
GTT	8
TAG	6, 10
TTA	9

Index of fixed-length words



ACGGCTAGTTAGAA*

GTTAG

TAG

AA*	13
ACG	1
AGA	11
AGT	7
CGG	2
CTA	5
GAA	12
GCT	4
GGC	3
GTT	8
TAG	6, 10
TTA	9

Index of fixed-length words

ACGGCTAGTTAGAA*

GTTAG

TAG

AA*	13
ACG	1
AGA	11
AGT	7
CGG	2
CTA	5
GAA	12
GCT	4
GGC	3
GTT	8
TAG	6, 10
TTA	9

Index of fixed-length words

ACGGCTAGTTAGAA*

CGA

AA*	13
ACG	1
AGA	11
AGT	7
CGG	2
CTA	5
GAA	12
GCT	4
GGC	3
GTT	8
TAG	6, 10
TTA	9

Index of fixed-length words

ACGGCTAGTTAGAA*

CGA

AA*	13
ACG	1
AGA	11
AGT	7
CGG	2
CTA	5
GAA	12
GCT	4
GGC	3
GTT	8
TAG	6, 10
TTA	9

Mathieu Giraud, Mikaël Salson, https://interstices.info/jcms/int_63223/les-sequenceurs-a-haut-debit

Suffix tree

ACGGCTAGTTAGAA$

ACGGCTAGTTAGAA$

$

A$

AA$

GAA$

AGAA$

TAGAA$

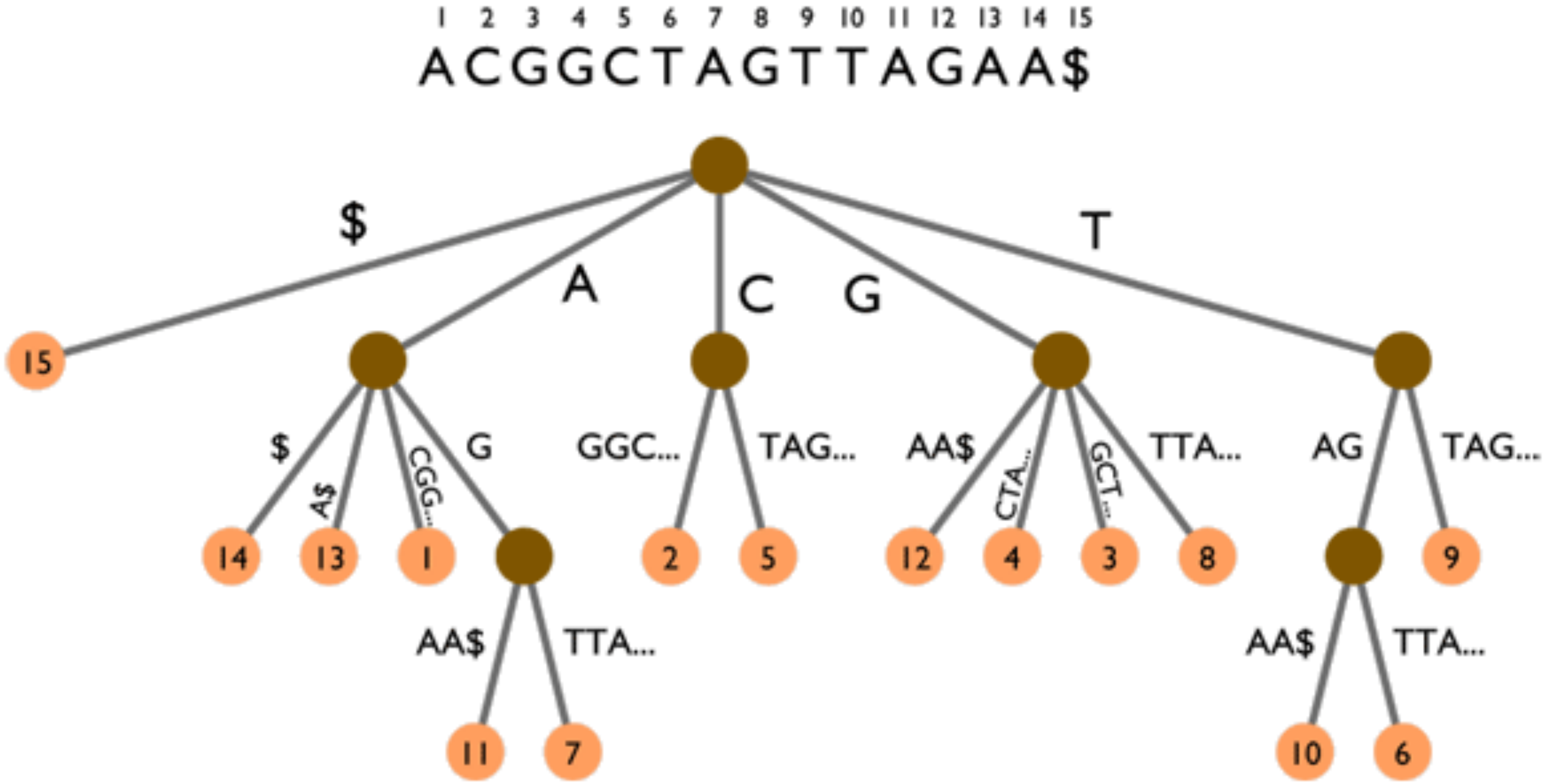
TTAGAA$

GTTAGAA$

AGTTAGAA$

...

Suffix tree



Mathieu Giraud, Mikaël Salson, https://interstices.info/jcms/int_63223/les-sequenceurs-a-haut-debit