Bit-encoding of canonical k-mers

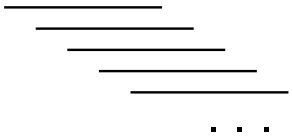
Roland Wittler

Bielefeld University

DSB 2022

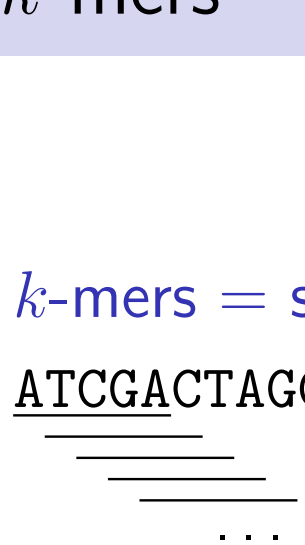
k-mers

k-mers = sequences of length k over alphabet $\{A, C, G, T\}$

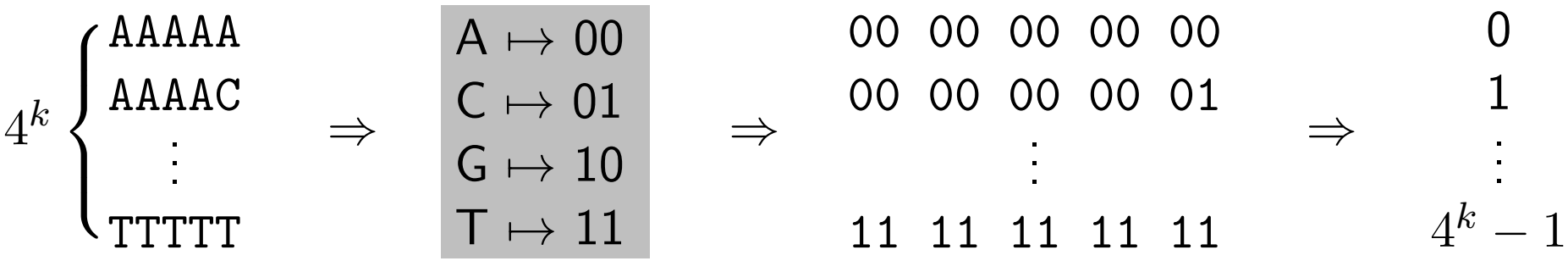


k-mers

k-mers = sequences of length k over alphabet $\{A, C, G, T\}$



Encoding



Canonical k-mers (k odd)

Canonical representative of k-mer $x = \min_{lex}\{x, \bar{x}\}$

AAAAA	AAAAC	...	~~ATTTT~~	...	TAAAA	...	~~TTTTG~~	~~TTTTT~~
~~TTTTT~~	~~GTTTT~~		AAAAT		~~TTTTA~~		CAAAA	AAAAA

Canonical k-mers (k odd)

Canonical representative of k-mer $x = \min_{lex}\{x, \bar{x}\}$

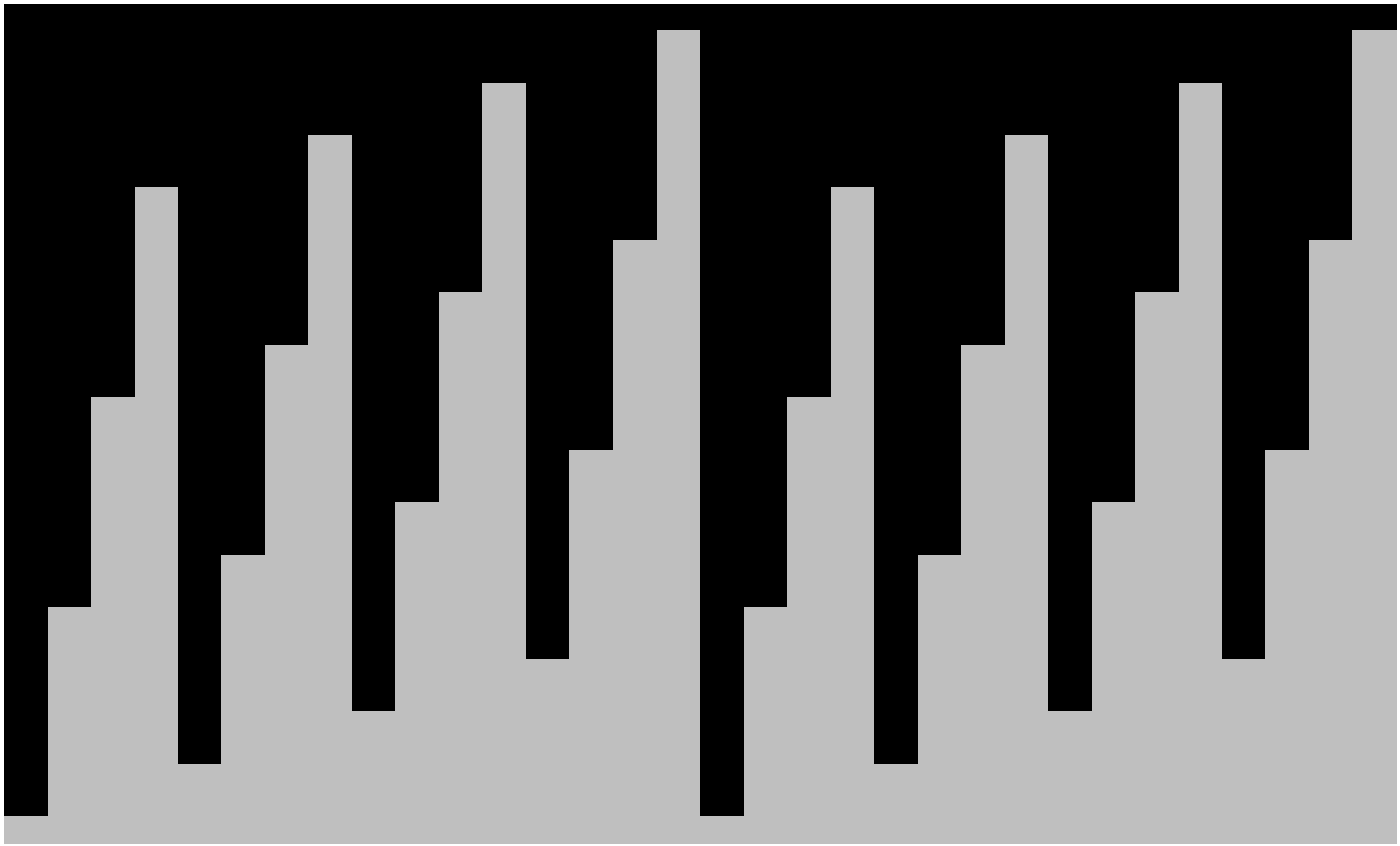
AAAAA	AAAAC	...	~~ATTTT~~	...	TAAAA	...	~~TTTTG~~	~~TTTTT~~
~~TTTTT~~	~~GTTTT~~		AAAAT		~~TTTTA~~		CAAAA	AAAAA

Encoding

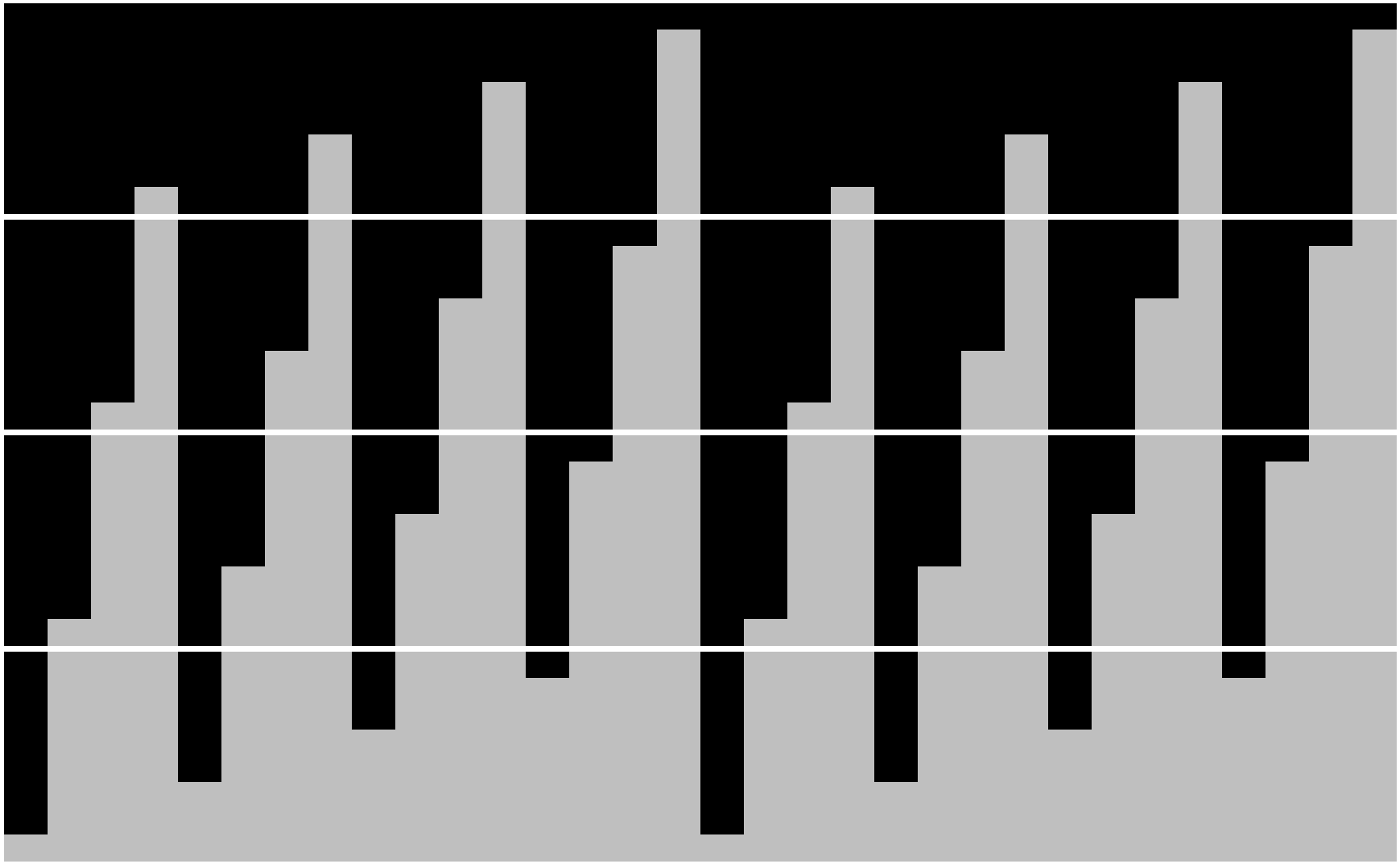
$\frac{4^k}{2}$ { AAAAA, AAAAC, ⋮, ~~TTTTT~~ } ⇒ { A ↦ 00, C ↦ 01, G ↦ 10, T ↦ 11 } ⇒

00	00	00	00	00	0
00	00	00	00	01	1
		⋮			⋮
~~11~~	~~11~~	~~11~~	~~11~~	~~11~~	~~$4^k - 1$~~

Canonical k-mers (k odd)



Canonical k-mers (k odd)

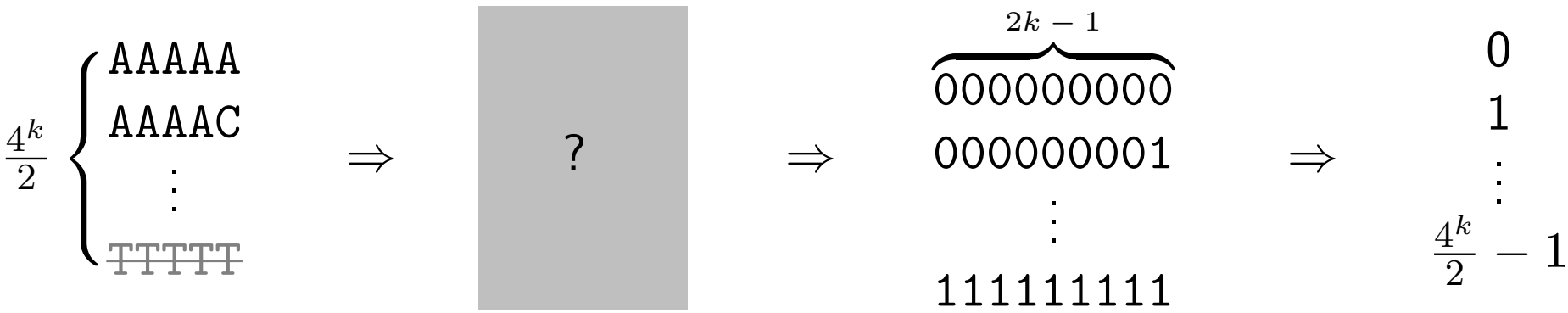


New encoding

canonical k-mers $= \frac{1}{2}$ # k-mers

- $2\,k - 1$ bits per k-mer

Encoding



New encoding

How to save one bit?

- What makes a k-mer canonical?

New encoding

How to save one bit?

- What makes a k-mer canonical?

Encode from the outside inwards

A G A T G A T

New encoding

How to save one bit?

- What makes a k-mer canonical?

Encode from the outside inwards

A G A T G A T

New encoding

How to save one bit?

- What makes a k-mer canonical?

Encode from the outside inwards

A G A T G A T

New encoding (k odd)

Encode from the outside inwards

A G A T G A T

11 11

unspecific pair

- A $\cdots$ T $\mapsto$ 11 $\cdots$ 11
- C $\cdots$ G $\mapsto$ 11 $\cdots$ 10
- G $\cdots$ C $\mapsto$ 11 $\cdots$ 01
- T $\cdots$ A $\mapsto$ 11 $\cdots$ 00

New encoding (k odd)

Encode from the outside inwards

A G A T G A T

11 10 1 11

unspecific pair

$$\texttt{A}\cdots\texttt{T} \mapsto \texttt{11}\cdots\texttt{11}$$
$$\texttt{C}\cdots\texttt{G} \mapsto \texttt{11}\cdots\texttt{10}$$
$$\texttt{G}\cdots\texttt{C} \mapsto \texttt{11}\cdots\texttt{01}$$
$$\texttt{T}\cdots\texttt{A} \mapsto \texttt{11}\cdots\texttt{00}$$

specifying case

$$\texttt{A}\cdots\texttt{A} \mapsto \texttt{00}\cdots\texttt{0}$$
$$\texttt{A}\cdots\texttt{C} \mapsto \texttt{00}\cdots\texttt{1}$$
$$\texttt{A}\cdots\texttt{G} \mapsto \texttt{01}\cdots\texttt{0}$$
$$\texttt{C}\cdots\texttt{A} \mapsto \texttt{01}\cdots\texttt{1}$$
$$\texttt{C}\cdots\texttt{C} \mapsto \texttt{10}\cdots\texttt{0}$$
$$\texttt{G}\cdots\texttt{A} \mapsto \texttt{10}\cdots\texttt{1}$$
$$\texttt{A} \mapsto \texttt{0}$$
$$\texttt{C} \mapsto \texttt{1}$$

New encoding (k odd)

Encode from the outside inwards

A G A T G A T

11 10 00 11 10 1 11

unspecific pair

$\mathtt{A}\cdots\mathtt{T} \mapsto \mathtt{11}\cdots\mathtt{11}$

$\mathtt{C}\cdots\mathtt{G} \mapsto \mathtt{11}\cdots\mathtt{10}$

$\mathtt{G}\cdots\mathtt{C} \mapsto \mathtt{11}\cdots\mathtt{01}$

$\mathtt{T}\cdots\mathtt{A} \mapsto \mathtt{11}\cdots\mathtt{00}$

specifying case

$\mathtt{A}\cdots\mathtt{A} \mapsto \mathtt{00}\cdots\mathtt{0}$

$\mathtt{A}\cdots\mathtt{C} \mapsto \mathtt{00}\cdots\mathtt{1}$

$\mathtt{A}\cdots\mathtt{G} \mapsto \mathtt{01}\cdots\mathtt{0}$

$\mathtt{C}\cdots\mathtt{A} \mapsto \mathtt{01}\cdots\mathtt{1}$

$\mathtt{C}\cdots\mathtt{C} \mapsto \mathtt{10}\cdots\mathtt{0}$

$\mathtt{G}\cdots\mathtt{A} \mapsto \mathtt{10}\cdots\mathtt{1}$

$\mathtt{A} \mapsto \mathtt{0}$

$\mathtt{C} \mapsto \mathtt{1}$

remainder

$\mathtt{A} \mapsto \mathtt{00}$

$\mathtt{C} \mapsto \mathtt{01}$

$\mathtt{G} \mapsto \mathtt{10}$

$\mathtt{T} \mapsto \mathtt{11}$

Even-length k-mers

Even-length k-mers

Palindromes

- $x = \overline{x} =$ canonical
- # palindromes $= 4^{k/2}$
- # non-palindromes $= 4^k - 4^{k/2}$
- # canonical k-mers $= \frac{1}{2}4^k - \frac{1}{2}4^{k/2} + 4^{k/2} \; > \mathbf{\frac{1}{2}4^k}$
- more than $2\,k - 1$ bits $\Rightarrow 2\,k$ bits

Even-length k-mers

Palindromes

- $x = \overline{x} =$ canonical
- # palindromes $= 4^{k/2}$
- # non-palindromes $= 4^k - 4^{k/2}$
- # canonical k-mers $= \frac{1}{2}4^k - \frac{1}{2}4^{k/2} + 4^{k/2} \; > \mathbf{\frac{1}{2}4^k}$
- more than $2\,k - 1$ bits $\Rightarrow 2\,k$ bits

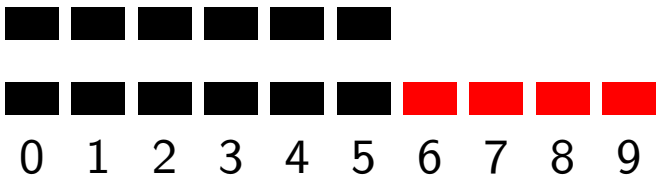
Tweak new encoding

- span integer range $0, \ldots,$ # canonical k-mers -1
- palindromes get highest ranks

Even-length k-mers

Palindromes get highest ranks?!

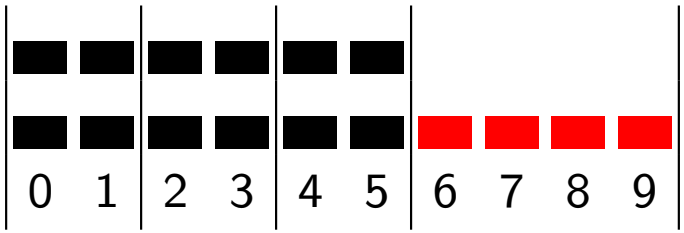
- non-palindromes: two k-mers $=$ one canonical k-mer
- palindrome: one k-mer $=$ one canonical k-mer



Even-length k-mers

Palindromes get highest ranks?!

- non-palindromes: two k-mers $=$ one canonical k-mer
- palindrome: one k-mer $=$ one canonical k-mer
- distribute canonical k-mers to buckets: heterogeneous bucket sizes



Even-length k-mers

00000000	AAAA
00000001	AAAC
00000010	AACA
00000011	AACC
⋮	⋮
01101100	TACA
01101100	TCAA
01101101	GACC
01101101	GCAC
01101110	CACG
01101110	CCAG
01101111	AACT
01101111	ACAT
01110000	TCCA
01110001	GCCC
01110010	CCCG
01110011	ACCT
01110100	TAGA
01110100	TGAA
01110101	GAGC
01110101	GGAC
01110110	CAGG
01110110	CGAG
01110111	AAGT
01110111	AGAT
01111000	TCGA
01111001	GCGC
01111010	CCGG
01111011	ACGT
01111100	TATA
01111101	GATC
01111110	CATG
01111111	AATT
10000000	TTAA
10000001	GTAC
10000010	CTAG
10000011	ATAT
10000100	TGCA
10000101	GGCC
10000110	CGCG
10000111	AGCT

Proof of concept

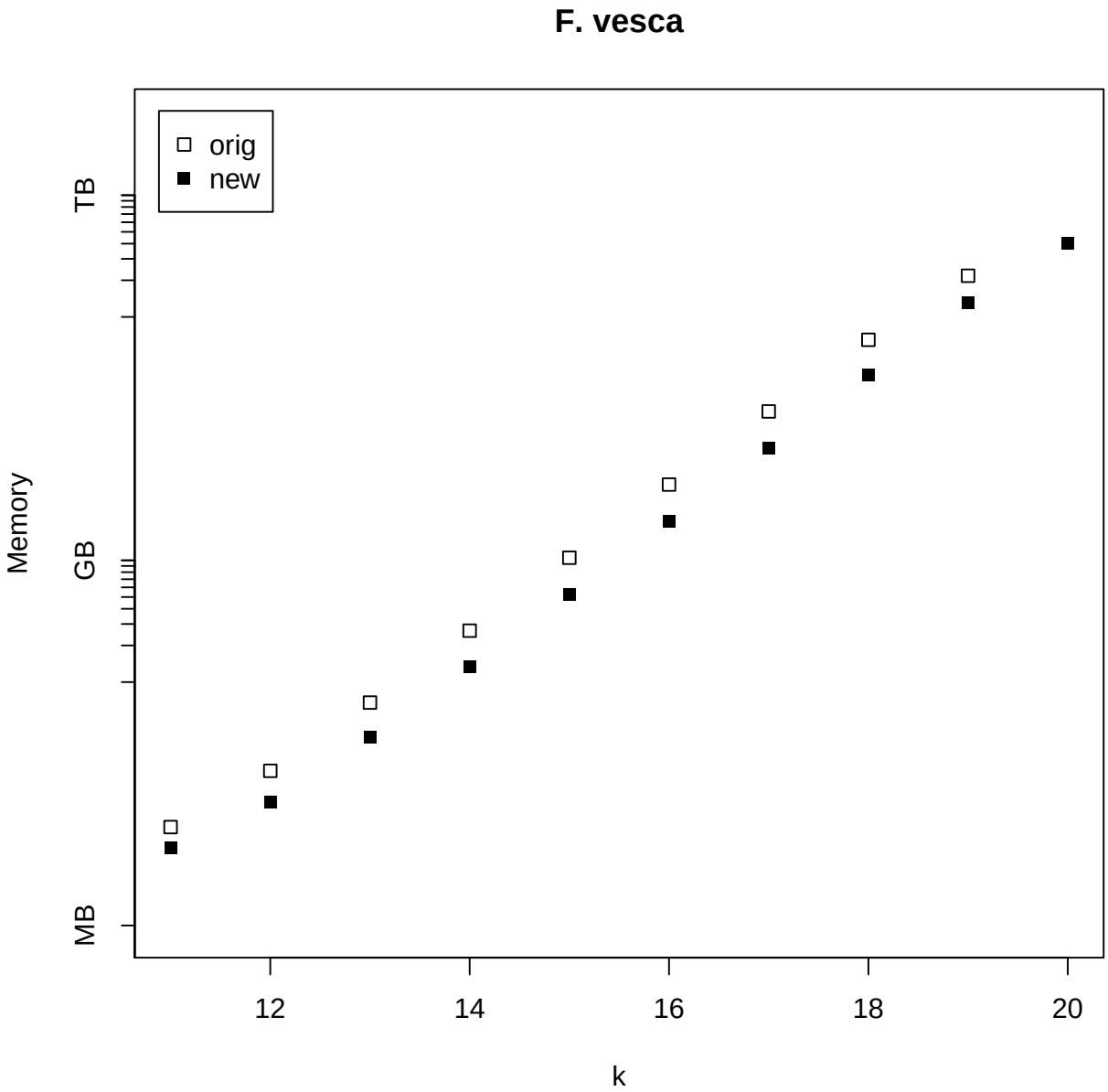
Frigate [Brihadiswaran & Jayasena, Int. Conf. Bioinf. & Biomed. Techn., 2021]

- k-mer counter
- in-memory: array of length 4^k
- multithreaded, lock-free array access

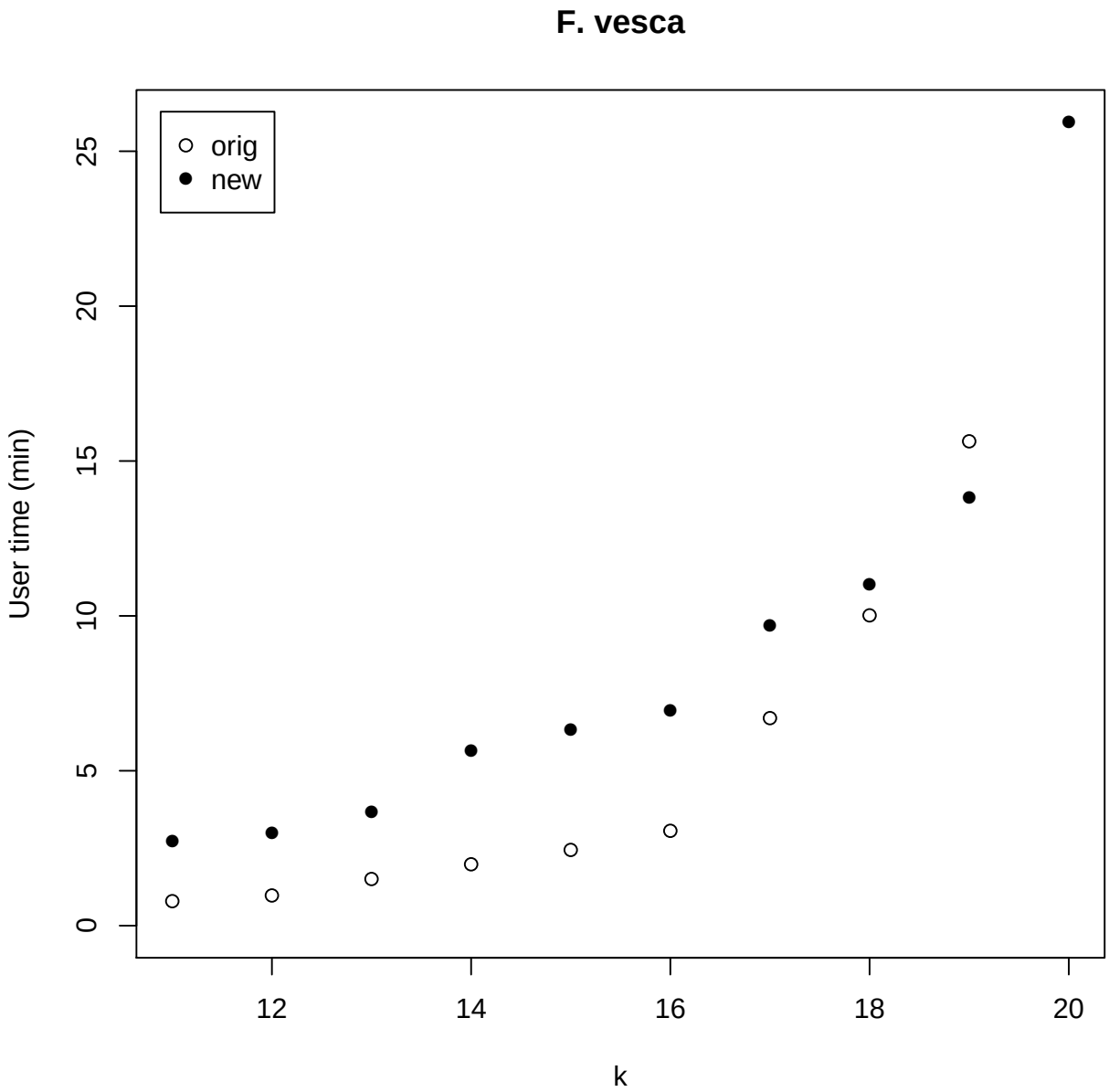
```
    minimum = kmer;
    if ( rc_kmer < kmer ) {
       minimum = rc_kmer;
    }
⇒   minimum = new_enc( kmer, rc_kmer );  //const. time

    pow(4, k_value)
⇒   pow(4, k_value) / 2   //k odd
```

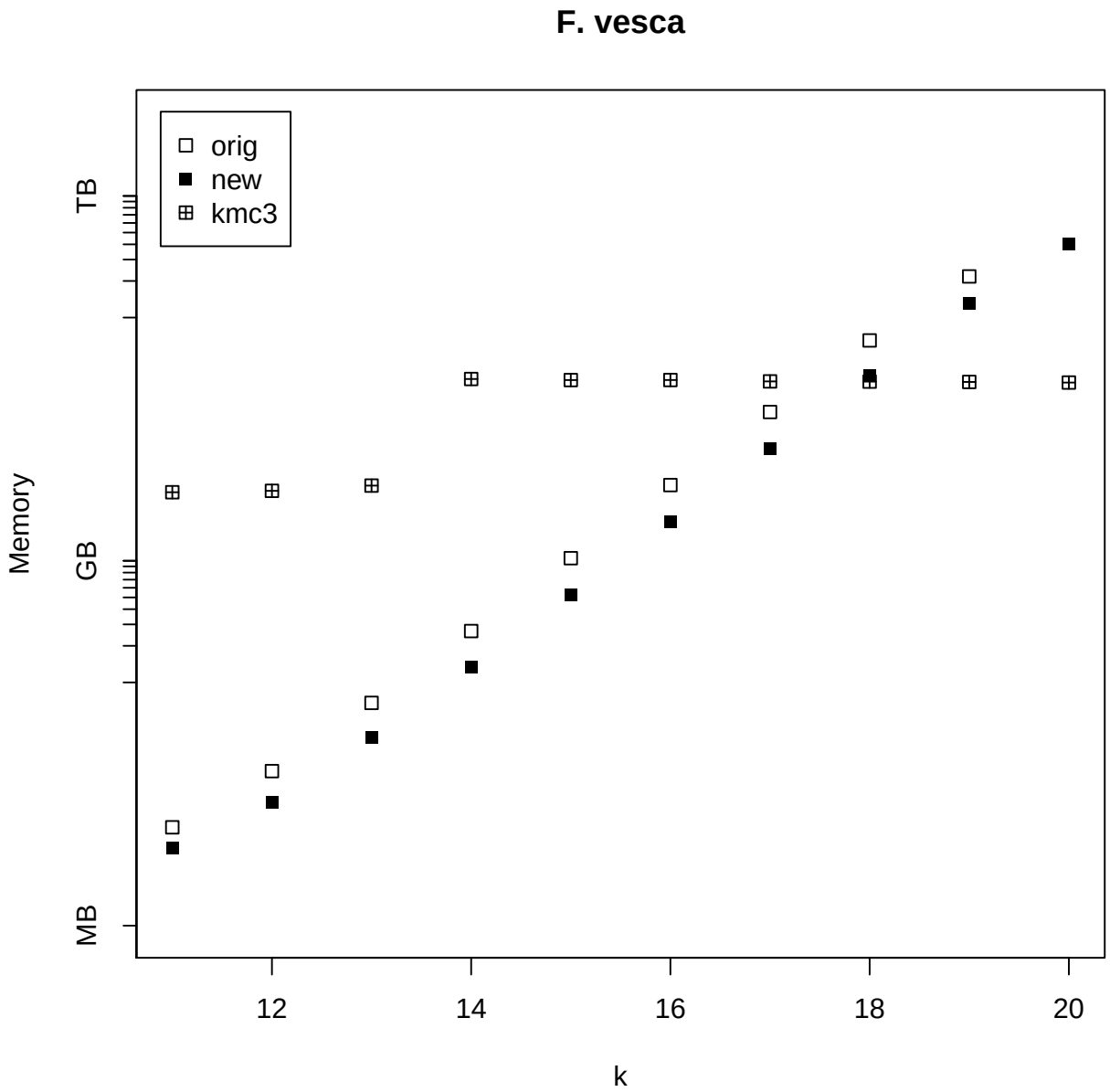
Proof of concept



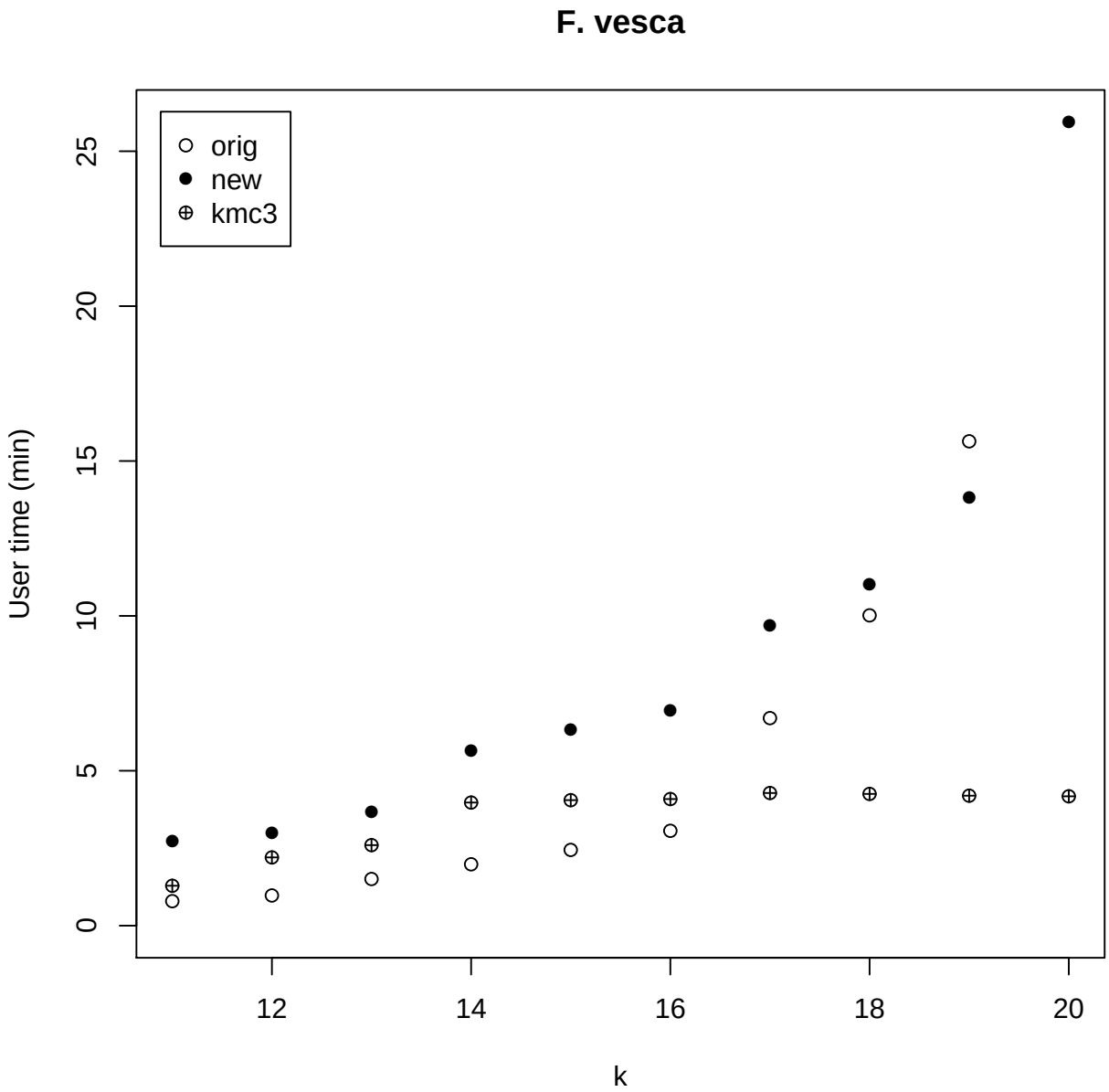
Proof of concept



Proof of concept



Proof of concept



Conclusion

Summary

- $2\,k - 1$ bit encoding of canonical k-mers
- including palindromes
- constant time transformation from $2\,k$ bit to $2\,k - 1$ bit

Conclusion

Summary

- $2\,k - 1$ bit encoding of canonical k-mers
- including palindromes
- constant time transformation from $2\,k$ bit to $2\,k - 1$ bit

Questions

- Other encodings?
- Useful?

Conclusion

Summary

- $2\,k - 1$ bit encoding of canonical k-mers
- including palindromes
- constant time transformation from $2\,k$ bit to $2\,k - 1$ bit

Questions

- Other encodings?
- Useful?

Thank you very much!

Outline

Appendix

New encoding (k odd)

Encode from the outside inwards

A G A T G A T

11 11

unspecific pair

- $\texttt{A}\cdots\texttt{T} \mapsto \texttt{11}\cdots\texttt{11}$
- $\texttt{C}\cdots\texttt{G} \mapsto \texttt{11}\cdots\texttt{10}$
- $\texttt{G}\cdots\texttt{C} \mapsto \texttt{11}\cdots\texttt{01}$
- $\texttt{T}\cdots\texttt{A} \mapsto \texttt{11}\cdots\texttt{00}$

New encoding (k odd)

Encode from the outside inwards

A G A T G A T

11 10 1 11

unspecific pair

$\texttt{A}\cdots\texttt{T} \mapsto \texttt{11}\cdots\texttt{11}$

$\texttt{C}\cdots\texttt{G} \mapsto \texttt{11}\cdots\texttt{10}$

$\texttt{G}\cdots\texttt{C} \mapsto \texttt{11}\cdots\texttt{01}$

$\texttt{T}\cdots\texttt{A} \mapsto \texttt{11}\cdots\texttt{00}$

specifying case

$\texttt{A}\cdots\texttt{A} \mapsto \texttt{00}\cdots\texttt{0}$

$\texttt{A}\cdots\texttt{C} \mapsto \texttt{00}\cdots\texttt{1}$

$\texttt{A}\cdots\texttt{G} \mapsto \texttt{01}\cdots\texttt{0}$

$\texttt{C}\cdots\texttt{A} \mapsto \texttt{01}\cdots\texttt{1}$

$\texttt{C}\cdots\texttt{C} \mapsto \texttt{10}\cdots\texttt{0}$

$\texttt{G}\cdots\texttt{A} \mapsto \texttt{10}\cdots\texttt{1}$

$\texttt{A} \mapsto \texttt{0}$

$\texttt{C} \mapsto \texttt{1}$

New encoding (k odd)

Encode from the outside inwards

A G A T G A T

11 10 00 11 10 1 11

unspecific pair

$\texttt{A}\cdots\texttt{T} \mapsto \texttt{11}\cdots\texttt{11}$

$\texttt{C}\cdots\texttt{G} \mapsto \texttt{11}\cdots\texttt{10}$

$\texttt{G}\cdots\texttt{C} \mapsto \texttt{11}\cdots\texttt{01}$

$\texttt{T}\cdots\texttt{A} \mapsto \texttt{11}\cdots\texttt{00}$

specifying case

$\texttt{A}\cdots\texttt{A} \mapsto \texttt{00}\cdots\texttt{0}$

$\texttt{A}\cdots\texttt{C} \mapsto \texttt{00}\cdots\texttt{1}$

$\texttt{A}\cdots\texttt{G} \mapsto \texttt{01}\cdots\texttt{0}$

$\texttt{C}\cdots\texttt{A} \mapsto \texttt{01}\cdots\texttt{1}$

$\texttt{C}\cdots\texttt{C} \mapsto \texttt{10}\cdots\texttt{0}$

$\texttt{G}\cdots\texttt{A} \mapsto \texttt{10}\cdots\texttt{1}$

$\texttt{A} \mapsto \texttt{0}$

$\texttt{C} \mapsto \texttt{1}$

remainder

$\texttt{A} \mapsto \texttt{00}$

$\texttt{C} \mapsto \texttt{01}$

$\texttt{G} \mapsto \texttt{10}$

$\texttt{T} \mapsto \texttt{11}$

New encoding (k odd)

Encode from the outside inwards

$$\mathtt{AGATGAT} = \overline{\mathtt{ATCATCT}}$$

11 10 00 11 10 1 11

unspecific pair

$\mathtt{A}\cdots\mathtt{T} \mapsto \mathtt{11}\cdots\mathtt{11}$

$\mathtt{C}\cdots\mathtt{G} \mapsto \mathtt{11}\cdots\mathtt{10}$

$\mathtt{G}\cdots\mathtt{C} \mapsto \mathtt{11}\cdots\mathtt{01}$

$\mathtt{T}\cdots\mathtt{A} \mapsto \mathtt{11}\cdots\mathtt{00}$

specifying case

$\mathtt{A}\cdots\mathtt{A} \mapsto \mathtt{00}\cdots\mathtt{0}$

$\mathtt{A}\cdots\mathtt{C} \mapsto \mathtt{00}\cdots\mathtt{1}$

$\mathtt{A}\cdots\mathtt{G} \mapsto \mathtt{01}\cdots\mathtt{0}$

$\mathtt{C}\cdots\mathtt{A} \mapsto \mathtt{01}\cdots\mathtt{1}$

$\mathtt{C}\cdots\mathtt{C} \mapsto \mathtt{10}\cdots\mathtt{0}$

$\mathtt{G}\cdots\mathtt{A} \mapsto \mathtt{10}\cdots\mathtt{1}$

$\mathtt{A} \mapsto \mathtt{0}$

$\mathtt{C} \mapsto \mathtt{1}$

remainder

$\mathtt{A} \mapsto \mathtt{00}$

$\mathtt{C} \mapsto \mathtt{01}$

$\mathtt{G} \mapsto \mathtt{10}$

$\mathtt{T} \mapsto \mathtt{11}$

New encoding (k odd)

Encode from the outside inwards

$$\underset{11\ 10\ 00\ 11\ 10\ \ 1\ \ 11}{\mathtt{A\,G\,A\,T\,G\,A\,T}} = \overline{\underset{11\qquad\qquad\qquad\qquad 11}{\mathtt{A\,T\,C\,A\,T\,C\,T}}}$$

unspecific pair

$\mathtt{A}\cdots\mathtt{T} \mapsto \mathtt{11}\cdots\mathtt{11}$

$\mathtt{C}\cdots\mathtt{G} \mapsto \mathtt{11}\cdots\mathtt{10}$

$\mathtt{G}\cdots\mathtt{C} \mapsto \mathtt{11}\cdots\mathtt{01}$

$\mathtt{T}\cdots\mathtt{A} \mapsto \mathtt{11}\cdots\mathtt{00}$

specifying case

$\mathtt{A}\cdots\mathtt{A} \mapsto \mathtt{00}\cdots\mathtt{0}$

$\mathtt{A}\cdots\mathtt{C} \mapsto \mathtt{00}\cdots\mathtt{1}$

$\mathtt{A}\cdots\mathtt{G} \mapsto \mathtt{01}\cdots\mathtt{0}$

$\mathtt{C}\cdots\mathtt{A} \mapsto \mathtt{01}\cdots\mathtt{1}$

$\mathtt{C}\cdots\mathtt{C} \mapsto \mathtt{10}\cdots\mathtt{0}$

$\mathtt{G}\cdots\mathtt{A} \mapsto \mathtt{10}\cdots\mathtt{1}$

$\mathtt{A} \mapsto \mathtt{0}$

$\mathtt{C} \mapsto \mathtt{1}$

remainder

$\mathtt{A} \mapsto \mathtt{00}$

$\mathtt{C} \mapsto \mathtt{01}$

$\mathtt{G} \mapsto \mathtt{10}$

$\mathtt{T} \mapsto \mathtt{11}$

New encoding (k odd)

Encode from the outside inwards

$$\mathtt{AGATGAT} = \overline{\mathtt{ATCATCT}}$$

11 10 00 11 10 1 11 | 11 10 1 11

unspecific pair

- $\mathtt{A}\cdots\mathtt{T} \mapsto 11\cdots11$
- $\mathtt{C}\cdots\mathtt{G} \mapsto 11\cdots10$
- $\mathtt{G}\cdots\mathtt{C} \mapsto 11\cdots01$
- $\mathtt{T}\cdots\mathtt{A} \mapsto 11\cdots00$

specifying case

- $\mathtt{A}\cdots\mathtt{A} \mapsto 00\cdots0$
- $\mathtt{T}\cdots\mathtt{T} \not\mapsto$
- $\mathtt{A}\cdots\mathtt{C} \mapsto 00\cdots1$
- $\mathtt{G}\cdots\mathtt{T} \not\mapsto$
- $\mathtt{A}\cdots\mathtt{G} \mapsto 01\cdots0$
- $\mathtt{C}\cdots\mathtt{T} \not\mapsto$
- $\mathtt{C}\cdots\mathtt{A} \mapsto 01\cdots1$
- $\mathtt{T}\cdots\mathtt{G} \not\mapsto$
- $\mathtt{C}\cdots\mathtt{C} \mapsto 10\cdots0$
- $\mathtt{G}\cdots\mathtt{G} \not\mapsto$
- $\mathtt{G}\cdots\mathtt{A} \mapsto 10\cdots1$
- $\mathtt{T}\cdots\mathtt{C} \not\mapsto$
- $\mathtt{A} \mapsto 0$
- $\mathtt{T} \not\mapsto$
- $\mathtt{C} \mapsto 1$
- $\mathtt{G} \not\mapsto$

remainder

- $\mathtt{A} \mapsto 00$
- $\mathtt{C} \mapsto 01$
- $\mathtt{G} \mapsto 10$
- $\mathtt{T} \mapsto 11$

New encoding (k odd)

Encode from the outside inwards

$$\mathtt{AGATGAT} = \overline{\mathtt{ATCATCT}}$$

11 10 00 11 10 1 11 | 11 10 00 11 10 1 11

unspecific pair

$\mathtt{A}\cdots\mathtt{T} \mapsto \mathtt{11}\cdots\mathtt{11}$

$\mathtt{C}\cdots\mathtt{G} \mapsto \mathtt{11}\cdots\mathtt{10}$

$\mathtt{G}\cdots\mathtt{C} \mapsto \mathtt{11}\cdots\mathtt{01}$

$\mathtt{T}\cdots\mathtt{A} \mapsto \mathtt{11}\cdots\mathtt{00}$

specifying case

$\mathtt{A}\cdots\mathtt{A} \mapsto \mathtt{00}\cdots\mathtt{0}$ ($\mathtt{T}\cdots\mathtt{T}$ ↝)

$\mathtt{A}\cdots\mathtt{C} \mapsto \mathtt{00}\cdots\mathtt{1}$ ($\mathtt{G}\cdots\mathtt{T}$ ↝)

$\mathtt{A}\cdots\mathtt{G} \mapsto \mathtt{01}\cdots\mathtt{0}$ ($\mathtt{C}\cdots\mathtt{T}$ ↝)

$\mathtt{C}\cdots\mathtt{A} \mapsto \mathtt{01}\cdots\mathtt{1}$ ($\mathtt{T}\cdots\mathtt{G}$ ↝)

$\mathtt{C}\cdots\mathtt{C} \mapsto \mathtt{10}\cdots\mathtt{0}$ ($\mathtt{G}\cdots\mathtt{G}$ ↝)

$\mathtt{G}\cdots\mathtt{A} \mapsto \mathtt{10}\cdots\mathtt{1}$ ($\mathtt{T}\cdots\mathtt{C}$ ↝)

$\mathtt{A} \mapsto \mathtt{0}$ ($\mathtt{T}$ ↝)

$\mathtt{C} \mapsto \mathtt{1}$ ($\mathtt{G}$ ↝)

reverse complement

remainder

$\mathtt{A} \mapsto \mathtt{00}$

$\mathtt{C} \mapsto \mathtt{01}$

$\mathtt{G} \mapsto \mathtt{10}$

$\mathtt{T} \mapsto \mathtt{11}$

New encoding (k odd)

Encode from the outside inwards

$$\mathtt{A\,G\,A\,T\,G\,A\,T} = \overline{\mathtt{A\,T\,C\,A\,T\,C\,T}}$$

11 10 00 11 10 1 11 | 11 10 00 11 10 1 11

unspecific pair

$\mathtt{A}\cdots\mathtt{T} \mapsto \mathtt{11}\cdots\mathtt{11}$

$\mathtt{C}\cdots\mathtt{G} \mapsto \mathtt{11}\cdots\mathtt{10}$

$\mathtt{G}\cdots\mathtt{C} \mapsto \mathtt{11}\cdots\mathtt{01}$

$\mathtt{T}\cdots\mathtt{A} \mapsto \mathtt{11}\cdots\mathtt{00}$

specifying case

$\mathtt{A}\cdots\mathtt{A} \mapsto \mathtt{00}\cdots\mathtt{0}$ ($\mathtt{T}\cdots\mathtt{T}$ ↝)

$\mathtt{A}\cdots\mathtt{C} \mapsto \mathtt{00}\cdots\mathtt{1}$ ($\mathtt{G}\cdots\mathtt{T}$ ↝)

$\mathtt{A}\cdots\mathtt{G} \mapsto \mathtt{01}\cdots\mathtt{0}$ ($\mathtt{C}\cdots\mathtt{T}$ ↝)

$\mathtt{C}\cdots\mathtt{A} \mapsto \mathtt{01}\cdots\mathtt{1}$ ($\mathtt{T}\cdots\mathtt{G}$ ↝)

$\mathtt{C}\cdots\mathtt{C} \mapsto \mathtt{10}\cdots\mathtt{0}$ ($\mathtt{G}\cdots\mathtt{G}$ ↝)

$\mathtt{G}\cdots\mathtt{A} \mapsto \mathtt{10}\cdots\mathtt{1}$ ($\mathtt{T}\cdots\mathtt{C}$ ↝)

$\mathtt{A} \mapsto \mathtt{0}$ ($\mathtt{T}$ ↝)

$\mathtt{C} \mapsto \mathtt{1}$ ($\mathtt{G}$ ↝)

reverse complement

remainder

$\mathtt{A} \mapsto \mathtt{00}$

$\mathtt{C} \mapsto \mathtt{01}$

$\mathtt{G} \mapsto \mathtt{10}$

$\mathtt{T} \mapsto \mathtt{11}$

Bijectiveness

- decoding

New encoding (k odd)

00000	AAA	10000	CAC
00001	AAC	10001	GAA
00010	ACA	10010	CCC
00011	ACC	10011	GCA
00100	AGA	10100	CGC
00101	AGC	10101	GGA
00110	ATA	10110	CTC
00111	ATC	10111	GTA
01000	AAG	11000	TAA
01001	CAA	11001	GAC
01010	ACG	11010	CAG
01011	CCA	11011	AAT
01100	AGG	11100	TCA
01101	CGA	11101	GCC
01110	ATG	11110	CCG
01111	CTA	11111	ACT

Efficient implementation

2-bit encoding: shift/update (const. time)

A G A T A T $\mapsto$ 00 10 00 11 00 11

G A T A T A $\mapsto$ 10 00 11 00 11 00

Efficient implementation

2-bit encoding: shift/update (const. time)

A G A T A T $\mapsto$ 00 10 00 11 00 11
G A T A T A $\mapsto$ 10 00 11 00 11 00

new encoding: shift/update (const. time) + shorten (const. time)

A G A T A T $\mapsto$ 00 10 00 11 00 11 $\mapsto$ 11 10 00 11 1 11
G A T A T A $\mapsto$ 10 00 11 00 11 01 $\mapsto$ 10 00 11 00 11 1

Efficient implementation

Shorten: 2-bit $\Rightarrow$ new encoding (const. time)

k-mer x	A C C C T C G T
$\overline{x}$	A C G A G G G T
2-bit: x	0001010111011011
$\overline{x}$	0001100010101011

_1111100111101011

Efficient implementation

Shorten: 2-bit $\Rightarrow$ new encoding (const. time)

k-mer x	A C C C T C G T
$\overline{x}$	A C G A G G G T
2-bit: x	0001010111011011
$\overline{x}$	0001100010101011
XOR	0000110101110000
count leading zeros	4

_1111100111O1011

Efficient implementation

Shorten: 2-bit $\Rightarrow$ new encoding (const. time)

k-mer x	A C C C T C G T
$\overline{x}$	A C G A G G G T
2-bit: x	0001010111011011
$\overline{x}$	0001100010101011
XOR	0000110101110000
count leading zeros	4
	0000000000001011
	_**1111**0101110**0000**
OR	_111101011101011
set specifying pos.	_111110011101011

Even-length k-mers

Encoding palindromes

A G A T C T

11 11 01 11

unspecific pair

$\texttt{A}\cdots\texttt{T} \mapsto \texttt{11}\cdots\texttt{11}$

$\texttt{C}\cdots\texttt{G} \mapsto \texttt{11}\cdots\texttt{10}$

$\texttt{G}\cdots\texttt{C} \mapsto \texttt{11}\cdots\texttt{01}$

$\texttt{T}\cdots\texttt{A} \mapsto \texttt{11}\cdots\texttt{00}$

specifying case

$\texttt{A}\cdots\texttt{A} \mapsto \texttt{00}\cdots\texttt{0}$

$\texttt{A}\cdots\texttt{C} \mapsto \texttt{00}\cdots\texttt{1}$

$\texttt{A}\cdots\texttt{G} \mapsto \texttt{01}\cdots\texttt{0}$

$\texttt{C}\cdots\texttt{A} \mapsto \texttt{01}\cdots\texttt{1}$

$\texttt{C}\cdots\texttt{C} \mapsto \texttt{10}\cdots\texttt{0}$

$\texttt{G}\cdots\texttt{A} \mapsto \texttt{10}\cdots\texttt{1}$

$\texttt{A} \mapsto \texttt{0}$

$\texttt{C} \mapsto \texttt{1}$

remainder

$\texttt{A} \mapsto \texttt{00}$

$\texttt{C} \mapsto \texttt{01}$

$\texttt{G} \mapsto \texttt{10}$

$\texttt{T} \mapsto \texttt{11}$

Even-length k-mers

Encoding palindromes

A G A T C T

11 11 01 11

unspecific pair

$\texttt{A}\cdots\texttt{T} \mapsto \texttt{11}\cdots\texttt{11}$

$\texttt{C}\cdots\texttt{G} \mapsto \texttt{11}\cdots\texttt{10}$

$\texttt{G}\cdots\texttt{C} \mapsto \texttt{11}\cdots\texttt{01}$

$\texttt{T}\cdots\texttt{A} \mapsto \texttt{11}\cdots\texttt{00}$

specifying case

$\texttt{A}\cdots\texttt{A} \mapsto \texttt{00}\cdots\texttt{0}$

$\texttt{A}\cdots\texttt{C} \mapsto \texttt{00}\cdots\texttt{1}$

$\texttt{A}\cdots\texttt{G} \mapsto \texttt{01}\cdots\texttt{0}$

$\texttt{C}\cdots\texttt{A} \mapsto \texttt{01}\cdots\texttt{1}$

$\texttt{C}\cdots\texttt{C} \mapsto \texttt{10}\cdots\texttt{0}$

$\texttt{G}\cdots\texttt{A} \mapsto \texttt{10}\cdots\texttt{1}$

$\texttt{A} \mapsto \texttt{0}$

$\texttt{C} \mapsto \texttt{1}$

palindrome identified

$\texttt{AT} \mapsto \texttt{011}\cdots\texttt{111}$

$\texttt{CG} \mapsto \texttt{011}\cdots\texttt{110}$

$\texttt{GC} \mapsto \texttt{100}\cdots\texttt{001}$

$\texttt{TA} \mapsto \texttt{100}\cdots\texttt{000}$

Even-length k-mers

Encoding palindromes

A G A T C T

0 11 11 111 01 11

unspecific pair

$\texttt{A}\cdots\texttt{T} \mapsto \texttt{11}\cdots\texttt{11}$

$\texttt{C}\cdots\texttt{G} \mapsto \texttt{11}\cdots\texttt{10}$

$\texttt{G}\cdots\texttt{C} \mapsto \texttt{11}\cdots\texttt{01}$

$\texttt{T}\cdots\texttt{A} \mapsto \texttt{11}\cdots\texttt{00}$

specifying case

$\texttt{A}\cdots\texttt{A} \mapsto \texttt{00}\cdots\texttt{0}$

$\texttt{A}\cdots\texttt{C} \mapsto \texttt{00}\cdots\texttt{1}$

$\texttt{A}\cdots\texttt{G} \mapsto \texttt{01}\cdots\texttt{0}$

$\texttt{C}\cdots\texttt{A} \mapsto \texttt{01}\cdots\texttt{1}$

$\texttt{C}\cdots\texttt{C} \mapsto \texttt{10}\cdots\texttt{0}$

$\texttt{G}\cdots\texttt{A} \mapsto \texttt{10}\cdots\texttt{1}$

$\texttt{A} \mapsto \texttt{0}$

$\texttt{C} \mapsto \texttt{1}$

palindrome identified

$\texttt{AT} \mapsto \texttt{011}\cdots\texttt{111}$

$\texttt{CG} \mapsto \texttt{011}\cdots\texttt{110}$

$\texttt{GC} \mapsto \texttt{100}\cdots\texttt{001}$

$\texttt{TA} \mapsto \texttt{100}\cdots\texttt{000}$

Even-length k-mers

Encoding palindromes

A G A T C T A G G C C T

0 11 11 111 01 11 1 00 00 001 01 11

unspecific pair

$\texttt{A}\cdots\texttt{T} \mapsto \texttt{11}\cdots\texttt{11}$

$\texttt{C}\cdots\texttt{G} \mapsto \texttt{11}\cdots\texttt{10}$

$\texttt{G}\cdots\texttt{C} \mapsto \texttt{11}\cdots\texttt{01}$

$\texttt{T}\cdots\texttt{A} \mapsto \texttt{11}\cdots\texttt{00}$

specifying case

$\texttt{A}\cdots\texttt{A} \mapsto \texttt{00}\cdots\texttt{0}$

$\texttt{A}\cdots\texttt{C} \mapsto \texttt{00}\cdots\texttt{1}$

$\texttt{A}\cdots\texttt{G} \mapsto \texttt{01}\cdots\texttt{0}$

$\texttt{C}\cdots\texttt{A} \mapsto \texttt{01}\cdots\texttt{1}$

$\texttt{C}\cdots\texttt{C} \mapsto \texttt{10}\cdots\texttt{0}$

$\texttt{G}\cdots\texttt{A} \mapsto \texttt{10}\cdots\texttt{1}$

$\texttt{A} \mapsto \texttt{0}$

$\texttt{C} \mapsto \texttt{1}$

palindrome identified

$\texttt{AT} \mapsto \texttt{011}\cdots\texttt{111}$

$\texttt{CG} \mapsto \texttt{011}\cdots\texttt{110}$

$\texttt{GC} \mapsto \texttt{100}\cdots\texttt{001}$

$\texttt{TA} \mapsto \texttt{100}\cdots\texttt{000}$