Fast Python sequence classification

Piotr Styczyński

MIM UW (ps386038)

2024-12-04

Seed and Extend approach

(k,w)-minimizers

GCTATTA → Hash → 0xA1
CTATTAC → Hash → 0x02
TATTACC → Hash → 0xFC
ATTACCT → Hash → 0x45

Find Min → 0x02 → CTATTAC

Seed: Minimizer

Spacing

GCTATTA → Pattern → G**X**TA**X**TA → Hash → 0xA1
GATACTA → Pattern → G**X**TA**X**TA → Hash → 0xA1
GACACTA → Pattern → G**X**CA**X**TA → Hash → 0xFC

Seeds: Spaced

Strobomers

GCTATTACC**TTAA**TGTGA**TGGA**C → Link → GCTATTAATGGA → Hash → 0xA1
GCTATCCCC**TTAA**TGGGA**TGGA**C → Link → GCTATTAATGGA → Hash → 0xA1
GC**CCTC**CCCT**TAAA**GGGA**TGGA**C → Link → CCTCTAAATGGA → Hash → 0xFC

Selected k-mers

Seeds: Strobomers