Pravdepodobnosť a E-value
(cvičenie)

Tomáš Vinař

25.10.2018

Hračkársky prípad

Dotaz: ATGCTCAAAC (dĺžka $m = 10$)

Databáza: (dĺžka $n = 300$)

accacttgcgcacgatttccagattcggtttccctgggcgcacgaagggc

ccacgaagcgGCTCAACccggagccttagttagaaggggggtctccgtca

agagagacggtaagttggagggtcactagcggtggactccgaatggaaac

actgaatagtggcagaacctaaacctcgttttggatttcctgaaaaaggc

aggcgctagaggaagaggcacgactgtgctagagataatcacttgtaaga

ccttggggatgggcttcgtatgcagaacgcgataaggtatcgaaaacgtg

Skórovacia schéma: zhoda $+1$, nezhoda -1, medzera -1

Lokálne zarovnanie so skóre $S = 6$

GCTCAAAC

GCTCA-AC

E-value: koľko očakávame lokálnych zarovnaní so skóre aspoň S v náhodnej databáze dĺžky n pri náhodnom dotaze dĺžky m

Náhodný dotaz a dazabáza

Dotaz: GTGCCTGCAG

Databáza:

cctctgatagccttgaaccgggcgagactcatacagacagtgctcctcgg
gcgataaccatgagatgacaggtccgatgctaatgttaacggacctacag
tgacatgttaaagtgtccattaagtttataccggaatcaacgagtgtccc
ccagcgcggcgaccgatggagccCCTGCAGgtatactcacttcaaggatt
accgctcggtgtaagttagtgttcagtcagactatactaagtattcagtt
atagagcgttagtaggtcgaccatgagcgggtaggGTGCCGAGatgtgaa

Počet výskytov: 2

Náhodný dotaz a dazabáza

Dotaz: `TCGACCGAAA`

Databáza:

`tactccattagggattataacgactaaagcccgtcgtggcgggatcactt`
`tgagattcaactttaacgcatcacagaggaatctgagacaaagcaaaacc`
`gatcataatgatcgatccaggtaataagtctccttgatggcgttagactg`
`gaaataacagttgacttccgactatagtttaatgaacgttcgtaattaga`
`cgatcgtgtaacttaaccaaaggctgcccccaaactagctgagtaatagc`
`tcgtcctgagcatgtaagagtcagcctccacggaacactgcaacgttctt`

Počet výskytov: 0

Náhodný dotaz a dazabáza

Dotaz: CCCGTCGTAG

Databáza:

cagcattagccccgttatttCGTCGTtctccaacgggtctgcctttctgg
aacgtggcgaaccttcacaggtcagtctgtcatcgcctgcgcttagagcg
gacggtactcgaaaggtcggttcagtgtggcgctggaaagaagaatagca
acacatgcactaatggaaggtcccagtggtgtgggacattctggaCCCGT
GTgtgccaacctatgtgagctccggcgttgactcggaggatgttaacaag
atcaagctgtaggcgacgatccccgccgggtttcctctactgcctcgagc

Počet výskytov: 2

Náhodný dotaz a dazabáza

Dotaz: AGGATGAGGA

Databáza:

ttatcgattctccggtgcgccagtacagcacaaggctcggatcctgtaaa
acactacaccttaaaaactaagtcAGGATGtgatctcccttaaGATGAGa
cagtctctaatgcggcgtagtgggaccctcgtgaccgagctaagcagttc
acaatgggcgctctgagcgattggctggagaccttgacttcccggtaggt
gtggtgttagttctgtgcccagagataaccatccaccgtaatggatctcg
taactttacGATGAAGAccggcatcatctcagttatatttctaggacggg

Počet výskytov: 3

Celkovo opakujeme 100 krát

$S = 6$, $m = 10$, $n = 300$, obsah GC 50%

Počet výskytov: 2, 0, 2, 3, 3, 1, 0, 1, 1, 1, 0, 0, 4, 2, 0, 1, 0, 1, 0, 0, 1, 0, 0, 4, 3, 1, 1, 0, 0, 0, 2, 3, 0, 0, 2, 1, 1, 1, 0, 0, 0, 0, 4, 1, 1, 0, 0, 1, 1, 1, 2, 2, 2, 0, 0, 2, 0, 1, 1, 0, 1, 2, 2, 1, 0, 0, 1, 1, 2, 0, 1, 0, 0, 1, 0, 3, 2, 0, 2, 2, 1, 0, 0, 2, 0, 0, 1, 2, 1, 1, 3, 2, 2, 1, 1, 0, 2, 0, 1, 3

Priemerný počet výskytov: 1.05

Keď celé opakujeme viackrát, dostávame hodnoty 0.99, 1.15, 1.02, 1.07, 0.98, ...

Správna hodnota E-value: 0.99