Algorithmen der Sequenzanalyse: Statistischer Vergleich von k-mer Häufigkeiten

22.10.2024

Prof. Michael Smmeth

Algorithm 1: HäufigeWörterNochSchneller($Text$,k)

```
for i = 0 to 4^k − 1 do
    Anzahl(i) ← 0
HäufigeWörter ← ∅
maxAnzahl ← 0
for i = 0 to |Text| − k do
    kmer ← Text(i, k)
    index ← kmerZuIndex(kmer)
    Anzahl(index) ← Anzahl(index) + 1
    if Anzahl(index) = maxAnzahl then
        HäufigeWörter ← HäufigeWörter ∪ {kmer}
    else if Anzahl(index) > maxAnzahl then
        HäufigeWörter ← {kmer}
        maxAnzahl ← Anzahl(index)
return HäufigeWörter
```

“Versteckte Nachrichten” im *oriC* von *Thermotoga petrophila*

Anwenden von *HäufigeWörter*(*Text*, *k*) auf

Text = *oriC* von *Thermotoga petrophila* (548 Nukleotide), und

k = 3, 4, ..., 9

ergibt:

k	**3**	**4**	**5**	**6**	**7**	**8**	**9**
häufigste Anzahl	24	11	6	6	5	5	5
häufigste(s) *k*-mer(e)	`ttt`	`tacc`	`gatca`	`tgatca`	`acctacc`	`acctacca`	`acctaccac`
	`att`		`tgatc`				

Frage: ist eine / welche dieser häufigsten k-mere verschiedener Längen ist "überraschend" häufig?

Statistik: wie oft ist *überraschend oft*?

k	3	4	5	6	7	8	9
häufigste Anzahl	24	11	6	6	5	5	5

je größer *k*, desto weniger Vorkommnisse (Kombinatorik)

Appproximation der Wahrscheinlichkeit p_f, dass ein *k*-mer in einem Text der Länge *N* über das Alphabet *A* mindestens *t*-mal* vorkommt:

$$p_f(N, A, k, t) \approx \frac{\binom{N-t(k-1)}{t}}{|A|^{(t-1)\,k}}$$

* nicht-überlappend

für DNA-Sequenzen mit $|A| = 4$

$$p_f(N, k, t) \approx \frac{\binom{N-t(k-1)}{t}}{4^{(t-1)\,k}}$$

z.B. $p_f(500, 9, 5) \sim 5 * 10^{-11}$ → “seeehr überraschend”

Statistik: wie oft ist *überraschend oft*?

Anwenden von *HäufigeWörter*(*Text*, *k*) für *k* = 3, 4, ..., 9 auf den *oriC*

- von *Thermotoga petrophila* ergibt:

k	3	4	5	6	7	8	9
häufigste Anzahl	24	11	6	6	5	5	**5**
$p_f(N, k, t)$	10^{-1}	$1* 10^{-2}$	$2* 10^{-2}$	10^{-5}	10^{-6}	10^{-8}	$\mathbf{10^{-11}}$

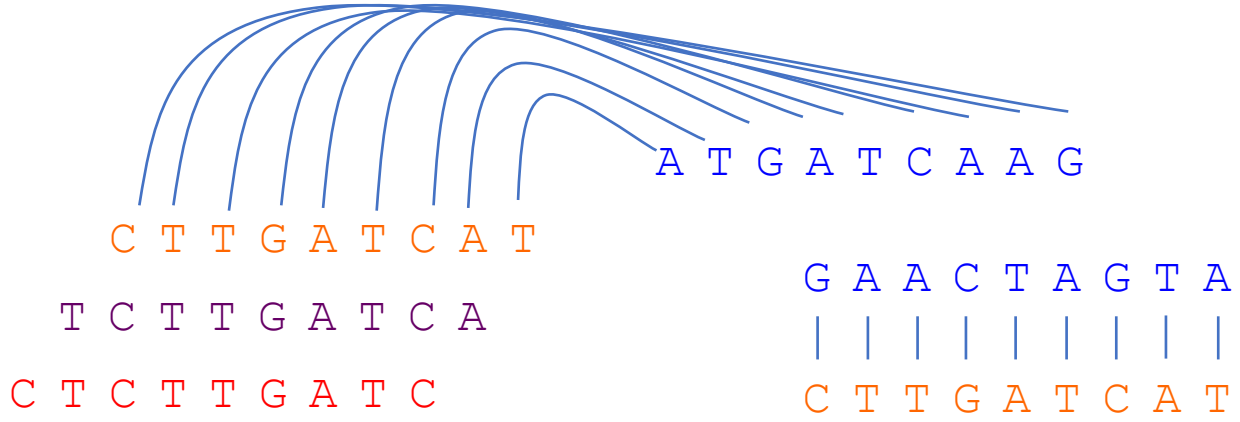
9-mere am "überraschensten",
Hypothese: DnaA-boxen sind Nonamere

- von *Vibrio cholerae* ergibt:

k	3	4	5	6	7	8	9
häufigste Anzahl	25	12	8	8	5	4	3
häufigste(s) *k*-mer(e)	tga	atga	gatca	tgatca	atgatca	atgatcaa	atgatcaag
			tgatc				cttgatcat
							tcttgatca
							ctcttgatc
$p_f(N, k, t)$	10^{-2}	10^{-3}	10^{-5}	10^{-9}	10^{-6}	10^{-5}	10^{-4}

Was bedeuten mehrere überraschend häufige *k*-mere?

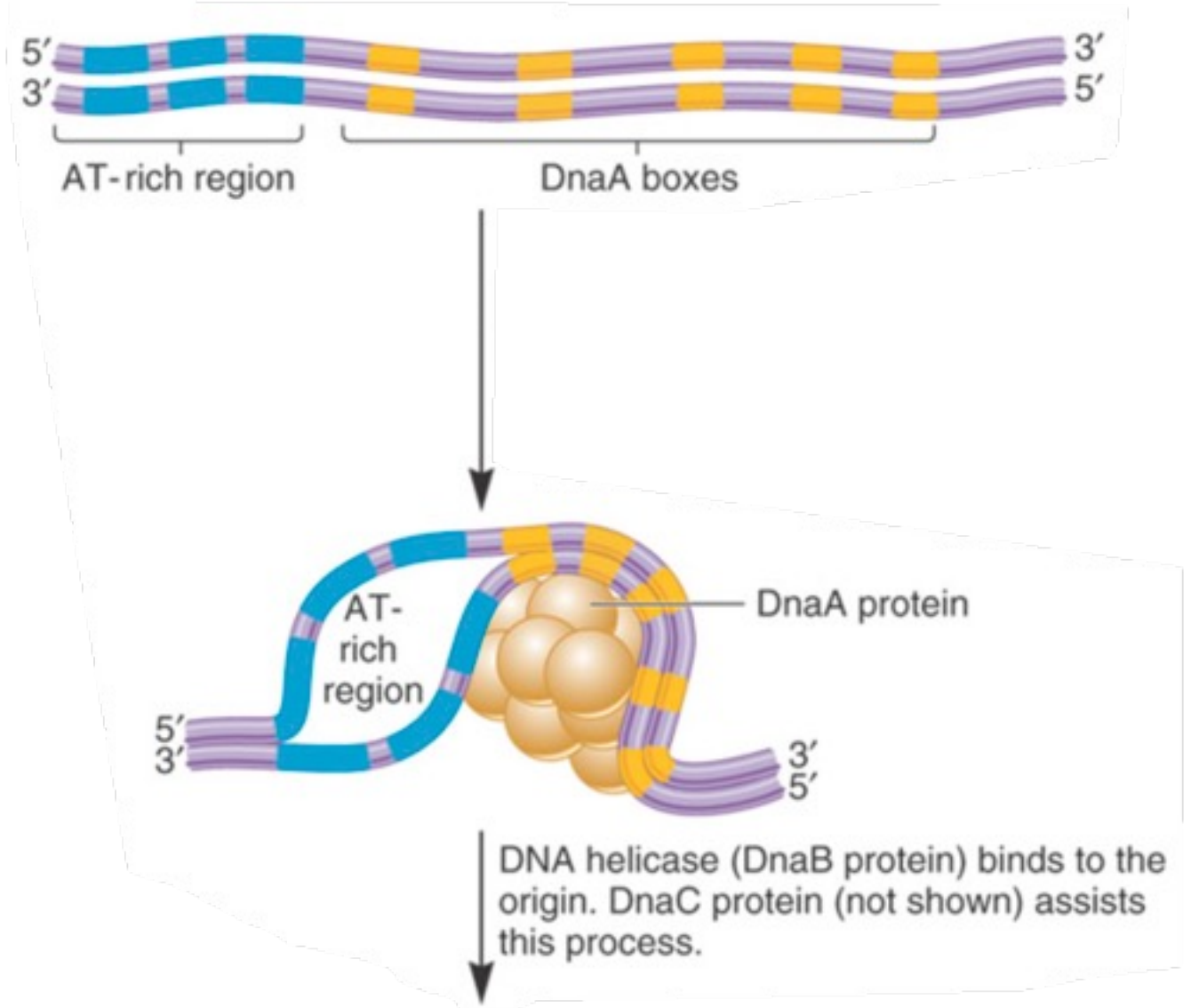
9-mere
atgatcaag
cttgatcat
tcttgatca
ctcttgatc



Im *oriC* von *Vibrio cholerae*:

atcaatgatcaacgtaagcttctaagcATGATCAAGgtgctcacacagtttatccacaac
ctgagtggatgacatcaagataggtcgttgtatctccttcctctcgtactctcatgacca
cggaaagATGATCAAGagaggatgatttcttggccatatcgcaatgaatacttgtgactt
gtgcttccaattgacatcttcagcgccatattgcgctggccaaggtgacggagcgggatt
acgaaagcatgatcatggctgttgttctgtttatcttgttttgactgagacttgttagga
tagacggtttttcatcactgactagccaaagccttactctgcctgacatcgaccgtaaat
tgataatgaatttacatgcttccgcgacgatttacctCTTGATCATcgatccgattgaag
atcttcaattgttaattctcttgcctcgactcatagccatgatgagctCTTGATCATgtt
tccttaaccctctatttttacggaagaATGATCAAGctgctgctCTTGATCATcgtttc

=



reverses Komplement muß bei häufigen Wörtern berücksichtigt werden!