SEQR

FAST K-MER COUNTING

JADWIGA SŁOWIK, PIOTR OLESIEJUK

K-MER COUNTING

- Analysing peptides sequences
- Counting occurrences of all k-length subsequences

seqR:: package for fast k-mer counting.

Algorithm description

Algorithm of seqR is based on hashtable that counts occurances of each k-mer in the sequence. The vector of elements is changed to integer representation for calculation of hash function and faster code execution. K-mers can be counted within specified alphabet, elements from outside of the alphabet are not counted. Implementation also allows to count gapped and positional k-mers. Positional ones are related to its position (column index in the given sequence matrix), so, for example k-mer "abc" that starts on the first position is a different k-mer than k-mer "abc" that starts on the second position. Original code was written in C++ and ported to R with Rcpp library. Parallelization was made with RcppParallel.

Function usage

count_kmers(s, d, alphabet, pos)

Params:

s	Sequence Matrix (non-empty), each row denotes a sequence. Sequences are represented as strings.
d	Integer Vector of gaps in k-mer. Length of this vector equals the length of k-mer + 1.
alphabet	Alphabet string Vector.
pos	Boolean value telling if k-mers are positional or not.

Returns:

Named Vector where names corresponds to k-mers and values to their counts.

Usage Example:

```
s  = matrix(data =c( 'A', 'B', 'C', 'A', 'C', 'G', 'B'), nrow = 1)
d = c(1) # 2-mers are counted with gap size = 1
alphabet  = c('A', 'B', 'C')
count_kmers(s,d,alphabet, pos)
pos = FALSE:
A.C  B.A  C.B  C.C  A.B  B.B  B.C  A.A  C.A
  1    1    1    1    0    0    0    0    0
pos = TRUE:
1_B.A  0_A.C  2_C.C  4_C.B
    1      1      1      1
```

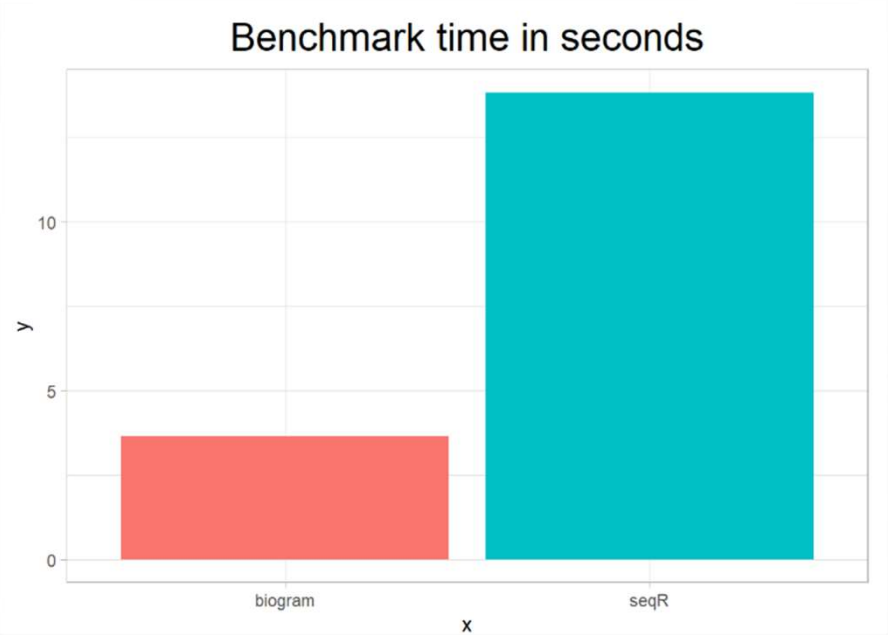
IMPLEMENTATION OVERVIEW

- Written in C++ with Rcpp library
- Parallelization with RcppParallel
- Hash table k-mer counting
- Integer sequence representation
- Gapped k-mers can be counted
- Counting over user defined alphabet

PACKAGE ELEMENTS

- Base Rcpp code
- Support for Windows, Linux and Mac OS
- Package Site
- Logo
- Cheetsheet
- Vignette

BENCHMARK



THANKS FOR YOUR ATTENTION