3 options for sequence generation (n-mers)

- Permutation with repetition (e.g. AAAAAAA, AAAAAAG….YYYYYYY) :

  -Number of entries = N^r

  - *N* = the amino acid alphabet
  - r = required number of letters in string (eg….the left side of HIP)
  - 20^7 = ***1.28 billion sequences***

- Sliding window of genome-protein sequences (method of choice)

  - Number of entries = $\sum_i^{n^{prot}} (l^{prot} - r)$
  - MUCH smaller, more efficient, realistic representation of sequence data, avoids generating too many false entries in the database.
  - For mouse: ~70k proteins (isomers,non reviewed, reviewed): ***5.6 million seqs.***

  AGTFDEWYYMSFGWEAANKLSDTRWSKDSCCFDSGYHJS

  AGTFDEW, GTFDEWY, TFDEWYY…..and so on….

3 options for sequence generation (n-mers)

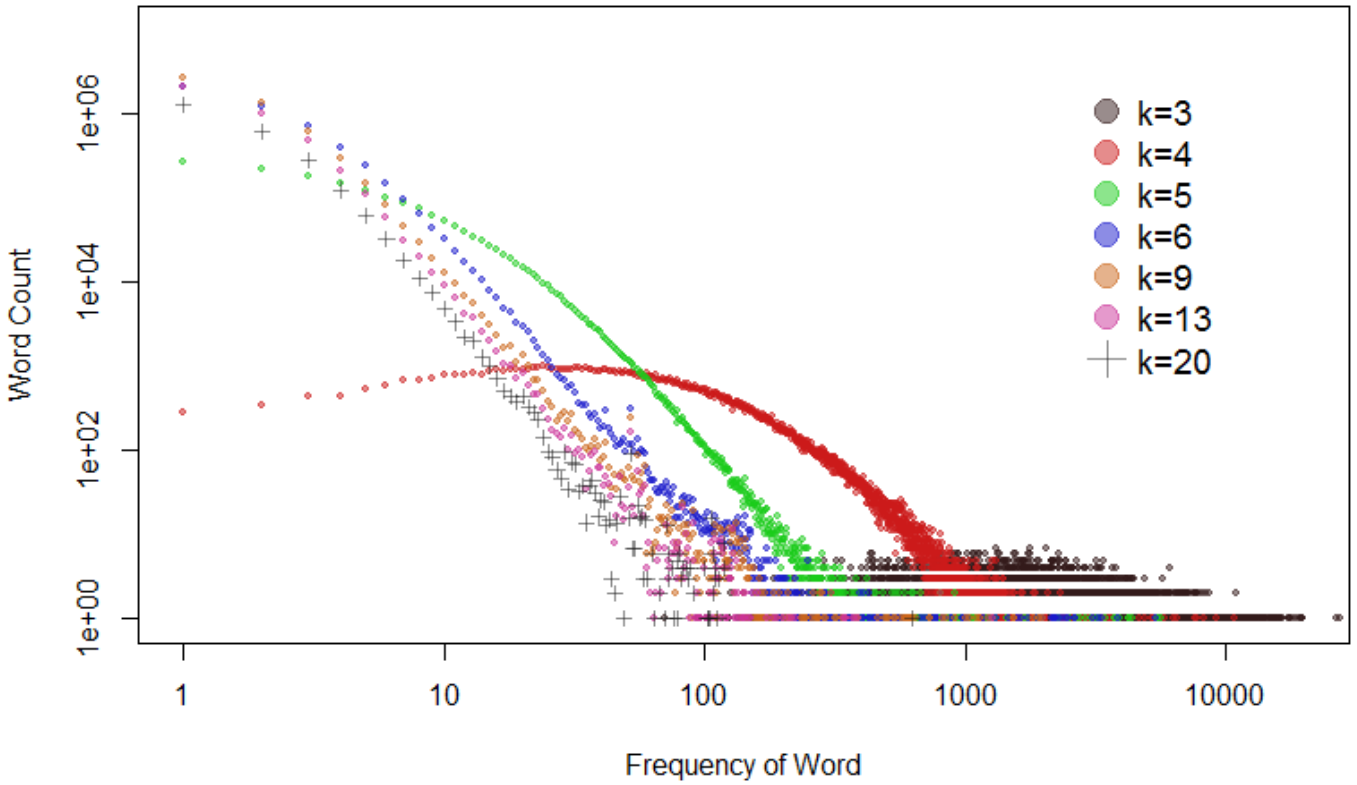
- Permutation around required motif (e.g. xxxDxRxG) :

  -Number of entries = $N^{(r-m)}$

  - *m*= the number of motif AAs in string
  - r = required number of random letters in string (eg….the left side of HIP)
  - 20^(7-3) = ***160000 sequences***
  - ***Requires biochemical elucidation of motif, or prior knowledge.***

  - ***If two motifs for both sides of chimeric are known, :*** $N_{Left}^{(r-m)} + N_{Right}^{(r-m)}$
  - ***Eg….*** 20^(7-3) + 20^(7-3) = 320000 Sequences.

HyPster C++ program for word counting in human protreome.

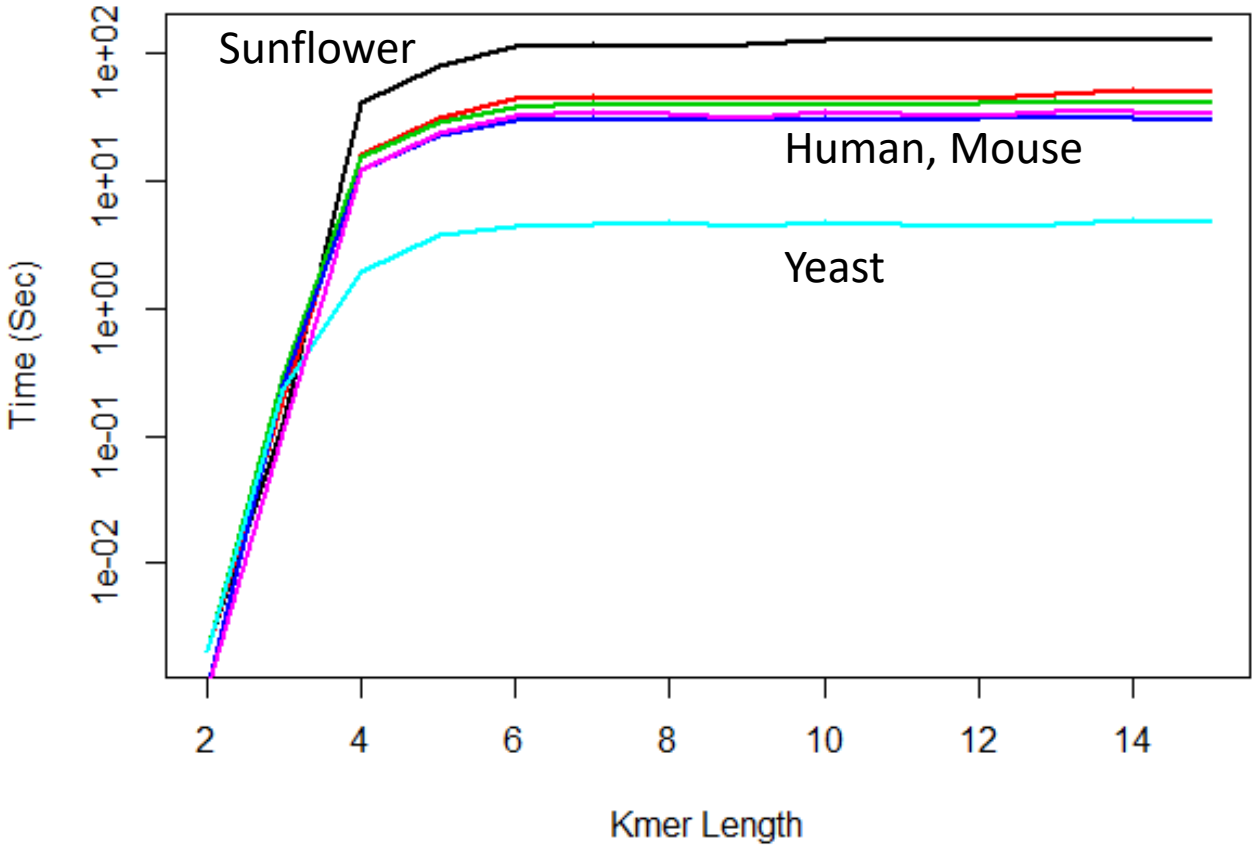


Sliding window algorithm.

Premise for probabilistic model on discrete data.

FAST.

HyPster C++ program for word counting in human protreome.



Sliding window algorithm.

Premise for probabilistic model on discrete data.

FAST