Kaiju Classification
Tutorial 9

Joshua Boisvert
Trung Hoang
Wenhan Tan

Existing Classifiers Problems

- Lack of sensitivity of overcoming evolutionary divergence (Large fractions of metagenomic reads remain unclassified)
- Slow computational methods with increasing volumes of microbial genome databases
    - New classifiers like Kraken depend on k-mers but only works best for samples have been previously sequenced and stored in the reference database (Also restricted at DNA level)
- Sampling bias (human microbiomes are over-represented in data since they are primary targets for microbial researches)
- Protein level classification is slow but increases accuracy and is more tolerant to sequencing errors (Degeneracy of the genetic code)
    - New classifiers like BastP are slow and report all alignments to the reference database, which need to be analysed further for taxonomic classification

What is Kaiju?

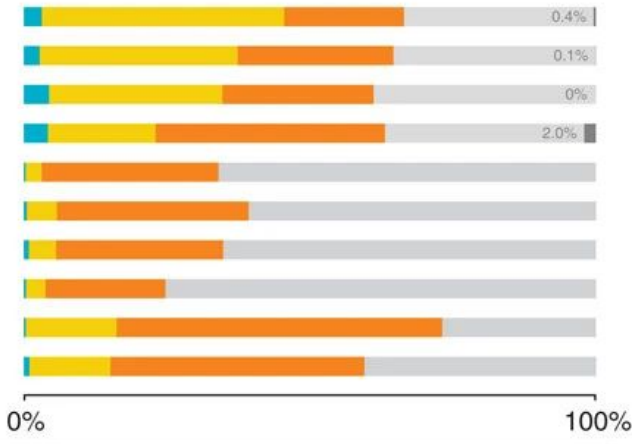
A protein-level metagenome classifier

- High sensitivity and precision
- Works with underrepresented genera in reference databases
- Uses Burrow-Wheeler transform (BWT, converts sequences into an easily searchable representation, which allows for exact string matching)
- Uses maximum exact matches (MEMs) and a lookup table of occurrence counts of each alphabet letter (FM-index, proposed by Ferragina and Manzini)
- Reads are assigned to a species or strain or to higher level nodes in the taxonomic tree
- Two modes: MEMs and Greedy (slower but larger search space)

Why MEMs?

- K-mer-based methods lack sensitivity and a big fraction of reads might remain unclassified
- MEMs is on protein level instead of nucleotide level to increase sensitivity
- Generally, MEMs on protein level comparison result in more classified reads

	MEM	Greedy	Kraken	Merged
Human vagina	58.6	63.3	45.5	66.4
Human saliva	56.1	61.8	37.4	64.6
Human gut	51.6	56.7	34.7	61.1
Cat gut	51.8	58.9	23.0	63.1
Lake	26.5	33.7	3.1	34.0
River plume	31.8	38.8	5.8	39.3
Baltic seawater	25.3	33.9	5.6	34.8
Desert soil	19.3	24.4	3.8	24.7
Bioreactor sediment	65.0	72.8	16.2	73.1
Bioreactor compost	50.3	58.5	15.1	59.5
Mean	43.6	50.3	19.0	52.1

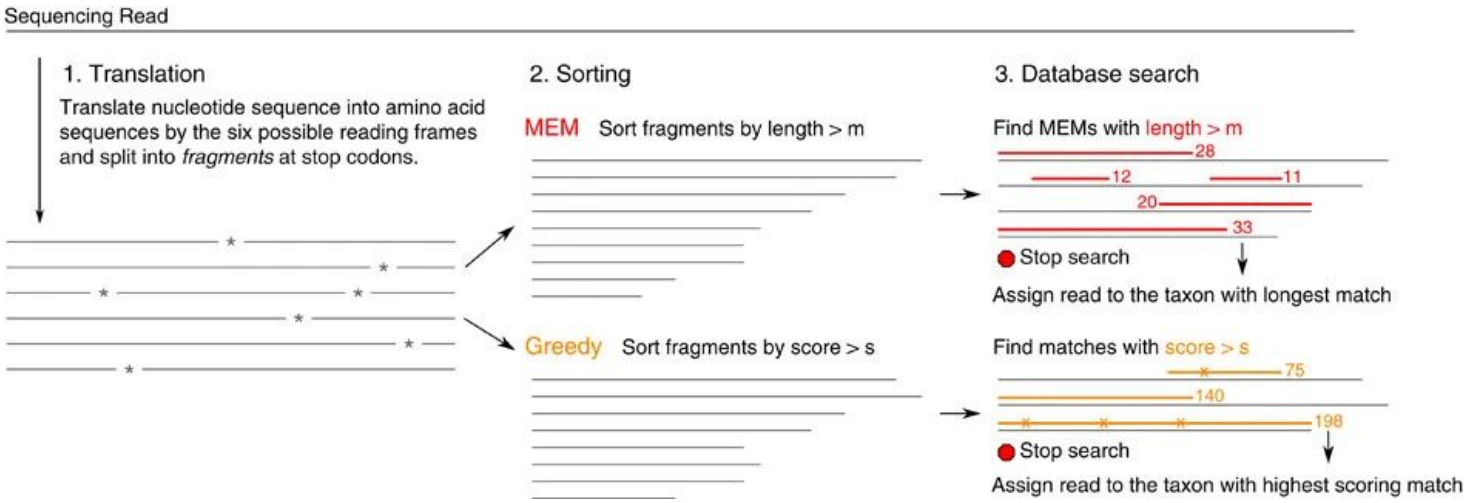


Columns - % of Classified Reads

Kraken - Blue
Greedy - Orange
MEM and Greedy Overlap - Yellow

% of reads mapped to respective host genomes

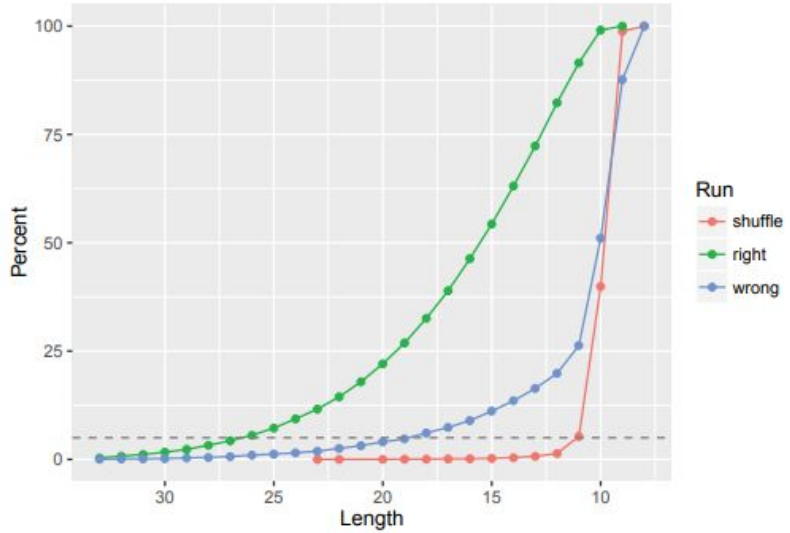
Algorithm



- Minimum required length **m**: 11
- Minimum required score **s**: 65
  - Because amino acid substitution in homologous sequences are non-uniform, speed-up can be gained by prioritizing the most likely substitutions with a total score called *BLOSUM62*

Determine Minimum Required Length (Same for minimum required score)

- Shuffled the microbial subset of NCBI NR protein database and search for MEMs between simulated reads and shuffled database
- 95% of data have length <= 11, 75% of wrong classification and 2% of correct classification have length <= 11



Demo

We perform both Greedy and Mem on 2 dataset:

- Refseq:
  - Completely assembled and annotated reference genomes of Archaea, Bacteria, and viruses from the NCBI RefSeq database.)
  - 50.9 M Sequences (31 GBs)
- Nr_euk:
  - Subset of NCBI BLAST database containing all proteins belonging to Archaea, Bacteria, Viruses, fungi and microbial eukaryotes
  - 178 M Sequences (83 GBs)

=> 4 combinations.

Repo: https://github.com/trung-hn/kaiju-classification

Results

- Greedy mode showed significantly higher precision (number of reads classified) than MEM mode for both databases
- There is an obvious trade off here between the % of the dataset that can be classified and the runtime of the algorithm
- As Greedy allows for mismatches it took substantially longer (several hours) than MEM for both reference databases
- We used the ‘kaiju2krona’ and kaiju2table’ scripts on the following github repository for visualization and analysis of the data, respectively: https://github.com/bioinformatics-centre/kaiju

Analysis of Kaiju Results (nr_euk db)

Run Mode	Greedy	MEM
% taxa group agreement with other run mode	97.46 %	96.24 %
Average # of reads per taxa group	3176	2913
% of taxa groups with more reads	30 %	70 %
# of reads unclassified	19630123 (63 %)	20548576 (66 %)

Analysis of Kaiju Results (refseq db)

Run Mode	Greedy	MEM
% taxa group agreement with other run mode	100 %	100 %
Average # of reads per taxa group	6442	4924
% of taxa groups with more reads	71 %	29 %
# of reads unclassified	23122332 (71 %)	24984233 (81 %)