Spaced Seeds Improve Metagenomic Classification

Karel Břinda, Maciej Sykulski, Gregory Kucherov
karel.brinda@univ-mlv.fr, macieksk@gmail.com, gregory.kucherov@univ-mlv.fr

Université Paris-Est Marne-la-Vallée

Metagenomics & NGS

Metagenomics is a powerful approach to study genetic content of environmental samples, which has been strongly promoted by NGS technologies.

For large-scale metagenomic projects alignment-free methods are used. To cope with massive data, recent tools (LMAT [1], Kraken [2]) rely on the analysis of **k-mers** (words of fixed size) shared between the **read** to be classified, and **reference genomes**.

Spaced Seeds

A **spaced seed** is a pattern over alphabet $A = \{\#, -\}$, where
$\#$ matching position, $-$ don't care position.

```
A A C A T T C T
  # # - # - #
A A C C T T C T
```

A seed acts as a mask for comparing short oligonucleotides. The number of #'s in a seed, called *weight*, defines the number k of matching nucleotides. In the above example k = 4, seed span = 6, and the matching (spaced) k-mer is **ACTC** signifying a **hit**.

A **coverage** is a number of aligned pairs covered by # from a spaced seed matches while sliding over an alignment (=4 above).

Contributions

- We showed that spaced seeds can improve success rate of binary classification of alignments into two categories, each defined by a specific mismatch rate. For example, in discriminating between alignments of length 100 with mismatch rate 0.2 and 0.3, a spaced seed of weight 16 achieves 63% of success while a contiguous seed of the same weight achieves only 40%.
- We demonstrated that spaced seeds allow for a better classification of NGS reads coming from a genome G between two other genomes G_1 and G_2 of the same genus.
- We analyzed how well different estimators (coverage/hit-number combined with spaced/contiguous seed) correlate with the alignment quality, also on real genomes.
- We demonstrated that spaced seeds can improve the sensitivity-selectivity trade-off in large-scale metagenomics experiments (SEED-KRAKEN panel).

References

[1] Ames, Sasha K., et al. Scalable metagenomic taxonomy classification using a reference genome database. In Bioinformatics 29.18 (2013): 2253-2260.

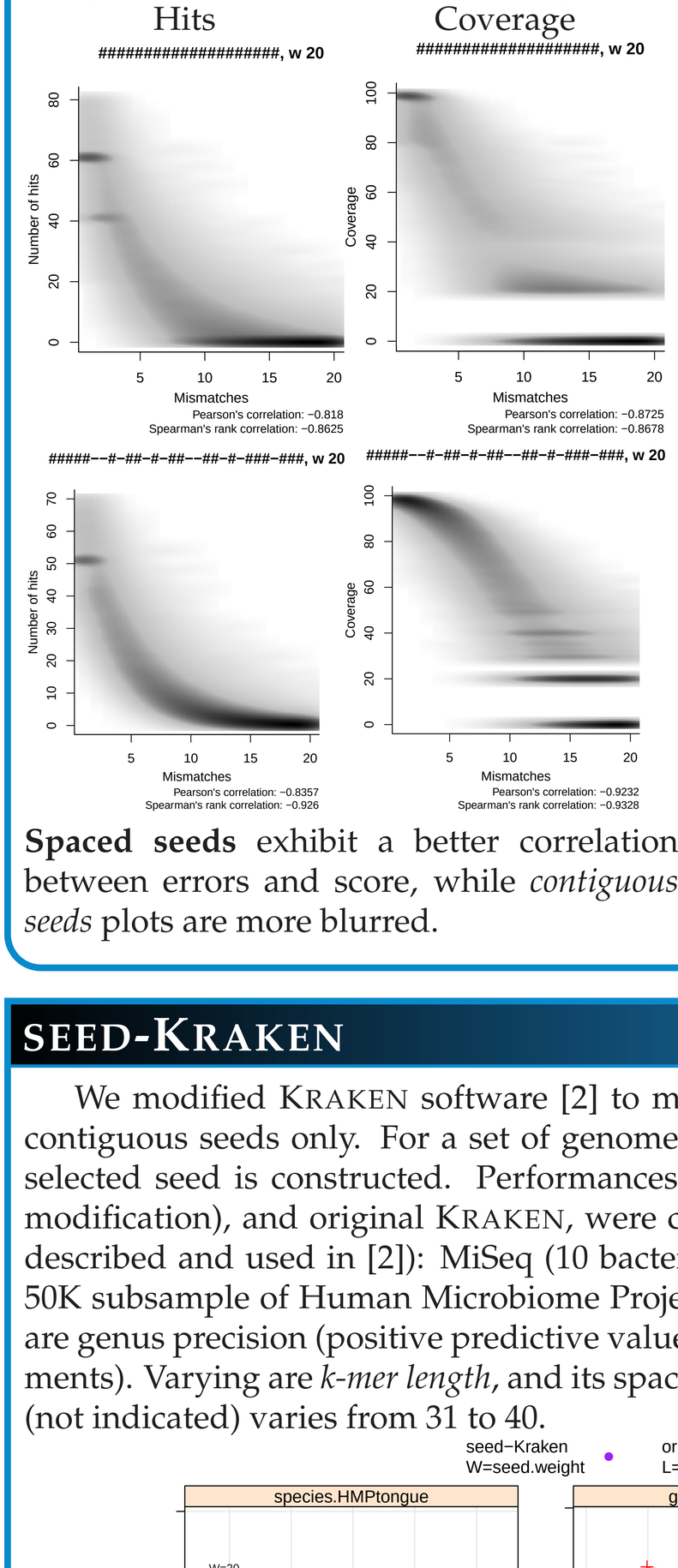
[2] Wood, D. E. and Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. In Genome Biol., 15(3), R46. (2014)

[3] Seed design has been done with IEDERA software `http://bioinfo.lifl.fr/yass/iedera`

`http://seed-kraken.readthedocs.org`
`http://github.com/gregorykucherov/spaced-seeds-for-metagenomics`
`http://arxiv.org/abs/1502.06256`

Scores on real genomes

We generated a set of ILLUMINA-like single-end reads: we've selected random substrings of *M.tuberculosis* genome of length $L = 100$ and introduced k mismatch errors, with k random between 1 and 20. For each read, we computed: **number of hits** and **coverage** to the genome under a given seed. A typical plot error vs score (seed weight 20).

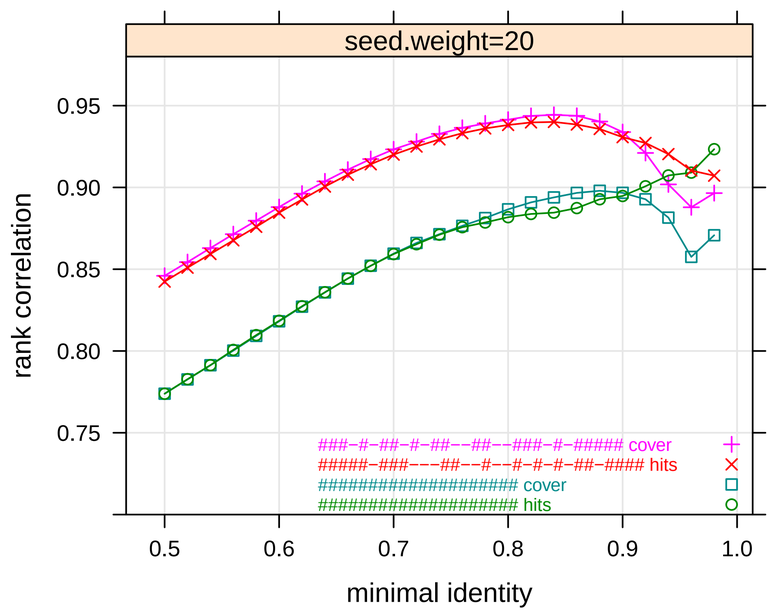


Spaced seeds exhibit a better correlation between errors and score, while *contiguous seeds* plots are more blurred.

Simulated Alignments

An accurate mapping of a read to a corresponding clade requires estimating its distances to each of the genomes.

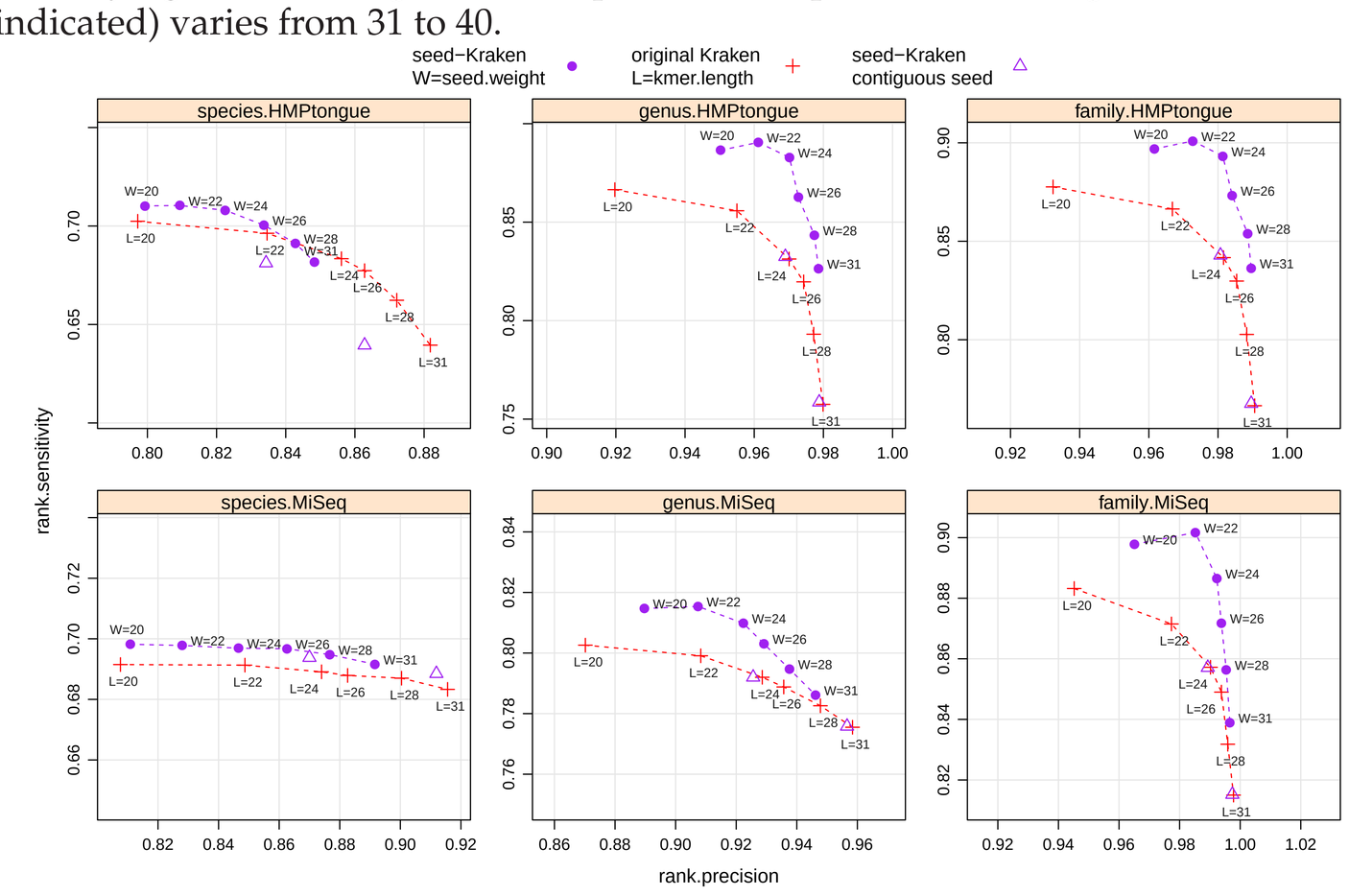
For a fixed minimal identity rate p_{id}, we randomly sampled gapless alignments of length 100 with identity rate from interval $[p_{id}..1]$, and collected pairs (number of mismatches, score), where 'score' stands for either **number of hits**, or **coverage** of a given seed. For these data, we plot Spearman's rank correlation.



In conclusion, spaced seeds provide a much better distance estimator for alignments whose score ranges over a large interval. For very high-scoring alignments ($> 95\%$ of identity), the hit number of contiguous seed becomes a better estimator.

seed-Kraken

We modified KRAKEN software [2] to make it work with spaced seeds rather than with contiguous seeds only. For a set of genomes, a database of spaced k-mers matching a user-selected seed is constructed. Performances of classification of SEED-KRAKEN (spaced seed modification), and original KRAKEN, were computed on simulated metagenomes (primarily described and used in [2]): MiSeq (10 bacterial genomes, average error rate), HiSeq, and on 50K subsample of Human Microbiome Project (HMP) Tongue Dorsum wgs sample. Charted are genus precision (positive predictive value) against genus sensitivity (rate of correct assignments). Varying are *k-mer length*, and its spaced seed equivalent *seed weight*, while the *seed span* (not indicated) varies from 31 to 40.



SEED-KRAKEN outperforms original KRAKEN in sensitivity/precision trade-off (ROC curve characteristics) at the classification levels of genus, and family.