IMPROVING MIN HASH FOR METAGENOMIC TAXONOMIC PROFILING

HOOMAN ZABETI[1], DAVID KOSLICKI[1*]

[1] *Mathematics Department, Oregon State University, Corvallis, OR.*

Abstract. Abstract here.

Keywords: *Min hash, k-mins sketch, metagenomics, taxonomic profiling, taxonomic classification, Jaccard index, containment.*

1. Introduction

(1) Min hash recently has been used to great success on biological data
(2) Mash, Titus' sourmash
(3) originally designed for sets of relatively similar size and appreciable intersection size
(4) metagenomic taxonomic profiling the setup is different: many relatively small database entries, one very large metagenomic sample, very small intersection sizes in general
(5) we modify the min hash paradigm to this particular situation so it can handle a sample of much greater size than the reference database entries.

2. Methods

Definitions, derivation of mathematical results here.

2.1. **Definitions.**

(1) Definitions of database entries, query sample, k-mer size, note size disparity
(2) define classic min hash (k-independent version and k-mins version)
(3) define the containment approach

2.2. **Min Hash via containment.**

2.3. **Time and space complexity.**

(1) Chernoff bound estimates
(2) comparison of number of hashes required for same accuracy
(3) time complexity
(4) space complexity (all with examples of the numbers in practice).

3. Results

In this section, we compare classic min hash to the proposed method.

Date: April 28, 2017.

* Corresponding Author: david.koslicki@math.oregonstate.edu.

3.1. **Synthetic data.** Here we illustrate the improved accuracy of containment min hashover classical min hash in estimating the Jaccard index. To that end, we generated two sets A and B consisting of 11-mers of two randomly generated strings w_A and w_B respectively on the alphabet $\{A, C, T, G\}$. We set $|w_A| = 10,000$ and $|w_B| = 15$ to simulate the situation of interest where one wishes to estimate the Jaccard index of two sets of very different size. We then appended a common string w_C of increasing length to each of w_A and w_B so that $\mathrm{Jac}(w_A w_C, w_B w_C)$ ranges between 0 and 1.

3.2. **Simulated biological data.**

3.3. **Real biological data.**

4. Discussion