SimkaMin: fast and resource frugal de novo comparative metagenomics

Gaëtan Benoit, Mahendra Mariadassou, Stéphane Robin, Sophie Schbath, Pierre Peterlongo, Claire Lemaitre

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Application Note

Application Note

SimkaMin: fast and resource frugal *de novo* comparative metagenomics

Gaetan Benoit [1], Mahendra Mariadassou [2], Stéphane Robin [3], Sophie Schbath [2], Pierre Peterlongo [1], and Claire Lemaitre [1,*]

[1]Univ Rennes, Inria, CNRS, IRISA, F-35000 Rennes, France
[2]MaIAGE, INRA, Université Paris-Saclay, 78350, Jouy-en-Josas, France
[3]UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris, France

*To whom correspondence should be addressed.

Associate Editor: XXXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: *De novo* comparative metagenomics is one of the most straightforward ways to analyze large sets of metagenomics data. Latest methods use the fraction of shared k-mers to estimate genomic similarity between read sets. However, those methods, while extremely efficient, are still limited by computational needs for practical usage outside of large computing facilities.

Results: We present SimkaMin, a quick comparative metagenomics tool with low disk and memory footprints, thanks to an efficient data subsampling scheme used to estimate Bray-Curtis and Jaccard dissimilarities. One billion metagenomic reads can be analyzed in less than 3 minutes, with tiny memory (1.09 GB) and disk (≈0.3 GB) requirements and without altering the quality of the downstream comparative analyses, making of SimkaMin a tool perfectly tailored for very large-scale metagenomic projects.

Availability: `https://github.com/GATB/simka`

Introduction

De novo comparative metagenomics is one of the most straightforward ways to analyze large sets of metagenomics data. It consists in estimating pairwise distances between samples, based only on their read content and using no *a priori* knowledge such as reference databases. In order to scale to very large datasets, all-versus-all sequence comparisons are avoided and the current fastest methods use long k-mers as comparison units, with $k > 20$. Typically, in the tool Simka (Benoit *et al.*, 2016), each sample is represented as a set of k-mers and various distances can be computed based on the amount and abundance of shared and specific k-mers between samples. Even if such methods are several orders of magnitude faster than alignment based methods, they are becoming too slow and resource-intensive for today's very-large metagenomic projects, such as the HMP (Lloyd-Price *et al.*, 2017) or TARA Oceans (Bork *et al.*, 2015), therefore limiting their practical use. We present here a new software called SimkaMin that drastically reduces the computational time by subsampling the k-mer space. Instead of considering the billions of k-mers typically present in metagenomic projects, genomic distances are estimated on randomly picked k-mers. k-mer subsampling is performed using the Minhash principle (Broder, 1997), that was introduced in the popular genome comparison tool Mash (Ondov *et al.*, 2016). Contrary to Mash, which can only compute presence/absence based distances, SimkaMin extends Minhash to the popular abundance-based Bray-Curtis dissimilarity. Taking k-mer abundances into account is indispensable when comparing metagenomics samples with highly variable species diversity profiles or in the general case of raw sequencing datasets where sequencing errors generate many low abundance k-mers.

Methods

Typically, a comparative tool would compute the Bray-Curtis dissimilarity $BC(S_1, S_2)$ from the counts of all the W k-mers occurring in a given pair of datasets S_1 and S_2. Here, the key idea is to compute instead an estimate $\widehat{BC}(S_1, S_2)$ of $BC(S_1, S_2)$ based on a reduced subset of w ($w \ll W$) randomly selected k-mers and their abundances in each dataset. In order to limit resource use, we never store the W k-mers but instead adapt the Minhash (Broder, 1997) principle to select the w k-mers on the fly.

Our approach is divided in two steps. First, it computes for each dataset a set of w k-mers together with their abundances, hereafter called a *sketch*. Then, it compares sketches in a pairwise fashion.

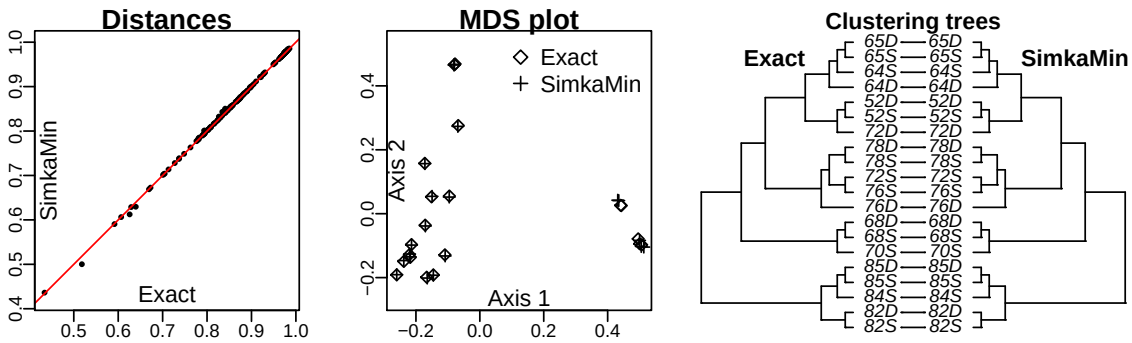

Fig. 1. Comparison of SimkaMin results (with $w = 10^6$ and default k-mer size (21)) with "exact" (i.e. without subsampling) results, in terms of raw distances (left), Multidimensional Scaling projection of the samples (middle) and clustering trees of the samples (right).

Sketching datasets A hash function h is used to transform each k-mer of a dataset into a 64 bits integer. These hashes are inserted in a sorted list L of fixed size w, that keeps only w distinct k-mers, those with the smallest hash values. The resulting w hashes represent a Minhash sketch m. Together with the list L, we maintain a structure that associates the abundance to each of the w k-mers present in L.

Distance computation For a pair of datasets (S_1, S_2), SimkaMin uses the associated sketches (m_1, m_2) to compute the Bray-Curtis dissimilarity $\widehat{BC}(S_1, S_2)$. The Bray-Curtis dissimilarity is computed using the counts of the w smallest (in the sense of their hash values) k-mers of the union $m_1 \cup m_2$. In practice the two sorted sketches are read in parallel, and the merge stops when w distinct hashes have been seen. Comparing two sketches has a $O(w)$ time complexity since sketches are already sorted.

Accuracy of the approximate BC dissimilarity

$\widehat{BC}(S_1, S_2)$ is a random estimate of $BC(S_1, S_2)$, as it depends on a random selection of k-mers. Standard sampling theory shows that both the bias and the standard deviation of $\widehat{BC}(S_1, S_2)$ decrease with the sketch size w. The bias decreases at speed $1/w$ and and the standard deviation at speed $1/\sqrt{w}$ and thus dominates the bias term for large w (see Supp. Mat.).

Usage

SimkaMin is implemented in C++ and is based on the GATB library (Drezen *et al.*, 2014). The main options are the k-mer size and the size w of the sketches SimkaMin can filter out k-mers seen only once in a dataset, as these k-mers are likely to contain sequencing errors. Output files are composed of the Bray-Curtis and the Jaccard dissimilarity matrices. The sketches of all input datasets are stored in a compressed binary file, that can be updated later with sketches of additional datasets thanks to a dedicated update python script.

Results

SimkaMin, was tested on a Tara oceans (Villar *et al.*, 2015) seawater metagenomic set composed of 20 datasets. Reference results are the ones without subsampling and were computed using Simka (Benoit *et al.*, 2016). Simka and SimkaMin were both parameterized to limit the computations to the first 50 millions reads of each dataset, thus analysing in total 1 billion reads (see Supp. Mat. for details and command lines).

Results presented in Table 1 show that a tiny amount of resources is sufficient to obtain highly accurate results. With default value ($w = 10^6$), compared to the so-called "exact" approach, SimkaMin provides results ≈11 times faster, with ≈3 times less RAM, and using ≈718 times less disk. Moreover, the bias due to subsampling is predictable and is sufficiently low not to alter the quality of the downstream analyses results, such as multidimensional scaling or sample clustering, as shown in Fig. 1. Similar experiments, performed on other metagenomic datasets from other biomes (soil and human gut), lead to similar conclusions (see Supp. Mat.).

w value	Time	Mem (GB)	Disk usage size (MB)	Correlation
10^4	2mn49	0.21	3	.9944
10^5	2mn49	0.99	31	.9976
10^6	3mn07	1.09	306	.9991
10^7	6mn59	12.12	3,051	.9994
unlimited	33mn12	3.71	22.10^4	1

Table 1. SimkaMin performances. w is the number of k-mers per sketch. The spearman correlation coefficient is computed between SimkaMin Bray-Curtis dissimilarities versus the non-sampled ones ("unlimited").

Acknowledgment

This work was supported by the French ANR-14-CE23-0001 Hydrogen Project. Computations have been performed thanks to the resources of the Genouest infrastructure.

References

Benoit,G., Peterlongo,P., Mariadassou,M., Drezen,E., Schbath,S., Lavenier,D. and Lemaitre,C. (2016) Multiple comparative metagenomics using multiset k-mer counting. *PeerJ Computer Science,* **2**, e94.

Bork,P., Bowler,C., De Vargas,C., Gorsky,G., Karsenti,E. and Wincker,P. (2015). Tara oceans studies plankton at planetary scale.

Broder,A.Z. (1997) On the resemblance and containment of documents. In *Compression and Complexity of Sequences 1997. Proceedings* pp. 21–29 IEEE.

Drezen,E., Rizk,G., Chikhi,R., Deltel,C., Lemaitre,C., Peterlongo,P. and Lavenier,D. (2014) GATB: genome assembly & analysis tool box. *Bioinformatics,* **30** (20), 2959–2961.

Lloyd-Price,J., Mahurkar,A., Rahnavard,G., Crabtree,J., Orvis,J. *et al.* (2017) Strains, functions and dynamics in the expanded human microbiome project. *Nature,* **550** (7674), 61.

Ondov,B.D., Treangen,T.J., Melsted,P., Mallonee,A.B., Bergman,N.H., Koren,S. and Phillippy,A.M. (2016) Mash: fast genome and metagenome distance estimation using minhash. *Genome Biol,* **17** (1), 132.

Villar,E., Farrant,G.K., Follows,M., Garczarek,L., Speich,S. *et al.* (2015) Environmental characteristics of agulhas rings affect interocean plankton transport. *Science,* **348** (6237), 1261447.