Pairwise and Incremental Multi-stage Alignment of Metagenomes: A New Proposal

Esteban Pérez-Wohlfeil, Oscar Torreno, and Oswaldo Trelles(✉)

Department of Computer Architecture, University of Malaga,
Boulevard Louis Pasteur 35, Malaga, Spain
ortrelles@uma.es

Abstract. Traditional comparisons between metagenomes are often performed using reference databases as intermediary templates from which to obtain distance metrics. However, in order to fully exploit the potential of the information contained within metagenomes, it becomes of interest to remove any intermediate agent that is prone to introduce errors or biased results. In this work, we perform an analysis over the state of the art methods and deduce that it is necessary to employ fine-grained methods in order to assess similarity between metagenomes. In addition, we propose our developed method for accurate and fast matching of reads.

Keywords: Next generation sequencing · Metagenome comparison · Read to read comparison · Distance between metagenomes

1 Introduction

New DNA acquisition methods have lowered sequencing costs up to a point where genomic research is producing an exponential growth in the number of sequenced samples, specially in the field of metagenomics. A metagenome is defined as an uncultured sample directly recovered from its original environment. Traditional metagenomic analysis comprises the comparison of one or multiple metagenomic samples against a reference database in order to find out the genomic composition, and to perform further analyses such as functional annotation. However, comparisons performed with a reference database do not fully exploit the data contained in metagenomes (e.g. species that are unknown or highly evolved that do not exactly correspond to those contained) and can lead to errors. For instance, reads belonging to samples that are not present in the reference database will most likely be aligned to their closest representatives. Furthermore, the results of the metagenomic comparison will probably change if the reference database is changed. Thus, the scenario of a direct comparison between metagenomes is gaining interest in order to produce results that are not biased by reference databases.

I. Rojas and F. Ortuño (Eds.): IWBBIO 2017, Part II, LNBI 10209, pp. 74–80, 2017.
DOI: 10.1007/978-3-319-56154-7_8

2 Background

Current methods developed in the field of metagenome-metagenome comparison are mostly based on analyzing the k-mer diversity of the samples (e.g. Compareads [1] or the more recent SIMKA [2]). Alignment-free methods have been reviewed several times (e.g. [3,4]) over the last decades, showing a persistent improvement in the characterization of distributions and handling of random matches. In this line, a variety of statistics have been proposed to compare genomic communities, such as the D2 statistic [5] or the Jaccard index [6]. In particular, most of the metagenomic comparison software use a variant –if not directly– of the Jaccard index (e.g. MASH [7] or SIMKA, along with other ecological distances). This index accounts for the number of shared k-mers between samples divided by the total number of different k-mers.

As shown in [8], analyses based on alignment-free methods can yield fairly accurate estimations of the similarity between samples. While such estimations are mostly used to perform classification of species, they do not serve to ensure relatedness of the reads that compose the metagenomes. In this line, COMMET [9] computes the number of shared reads (instead of k-mers) to produce a read-level similarity. However, the results of COMMET are very sensitive to the used parameters, and still do not show insight on the true correspondence between reads. Figure 1 shows an example where an incorrect classification of reads takes place due to the lack of inter alignments between shared k-mers. On the other hand, methods based on ungapped alignments such as BOWTIE [10] can not model evolutionary events such as insertions or deletions, which comprise a recurrent scenario in metagenomic studies. Typical gapped alignment approaches (e.g. BLAST [11]) often require large execution times and are not specifically designed to align reads to reads.

In this work, we show that the current software approaches for metagenomes comparison are coarse-grained, and that an exhaustive, gapped and fast alignment is required to improve both the assessment of similarity in metagenomes and the execution times.

3 Methods

The proposed method "IMSAME" (Incremental Multi Staged Alignment of MEtagenomes) performs an incremental alignment procedure, illustrated as follows:

1. A hash table of 12-mers is computed for the reference metagenome. The k-mer size is fixed at 12 nucleotides for three reasons: (a) providing sensitivity while retaining robustness [12]; (b) enabling the use of the algorithm without needing to parameterize the k-mer size, which often leads to strongly different results when changed; and (c) avoid the loss of candidate gapped alignments due to seed size to maximize results. In particular, the imposition of a fixed k-mer size avoids including parameters that are not intrinsically related to

the problem (i.e. similarity between samples) that is being solved. Thus, 12 was chosen as a fixed, highly-sensitive but still robust k-mer size.
2. After the hash table for the reference metagenome is computed, the query metagenome is then loaded and distributed to n threads. The algorithm is capable of working with any number of threads, from 1 to as many as the system allows to, thus enabling a massively parallel computation and large reduction of execution times compared to traditional software.
3. Each thread follows a multi-staged alignment step to compute gapped alignments:
   (a) Firstly, the matching 12-mer words between query and reference (hits) are computed.
   (b) A fast, first-approach ungapped alignment is performed for every hit. This is performed by linearly extending the hit in both forward and backward directions and keeping the starting and ending positions that yield the highest score. In this sense, the ungapped alignment continues until a negative score is reached, but the highest one is reported as the final alignment. Every ungapped alignment with an expected value less than the given threshold (typically, near zero, being default 10^{-5}) is considered as candidate for an exhaustive gapped alignment.
   (c) If such ungapped alignment exists, then the gapped alignment is computed using the Needleman-Wunsch global alignment algorithm between the reads that share the particular alignment. To speed up the Needleman-Wunsch algorithm computation time, a heuristic approach is used to enable gap insertion. The heuristic approach uses the stored maximum scores in the given row and column to insert gaps in case the diagonal score decays.

If the alignment produced in the multi-staged alignment step yields a high percentage of identity and coverage[1] (default 80% for both), the pair of reads are considered to be similar (and thus, shared between samples). In order to compute a global similarity measure between metagenomes, two approaches are considered: (1) the number of shared reads is divided by the total number of reads in the two samples, i.e. the Jaccard-index is computed at read level as shown in Eq. 1:

$$J(a,b) = \frac{r_a \cap r_b}{r_a \cup r_b} \tag{1}$$

where a and b are the metagenomes being compared, r_a and r_b are the reads contained in a and b respectively, and $J(a,b)$ is the Jaccard-index at read level. In the second approach (2) the percentage of reads from a contained in b is calculated to indicate the proportion of a that is contained in b. The resulting alignments are optionally written to disk to enable manual verification.

[1] Considering coverage as the length of the alignment divided by the length of the query read.

```
> Sample Read 1
CCGATTGCGAAGGCAGCCTGCTA{1}AGCTGCAACTGACATTGAGGCTCGAAAGTG{1}
TGGGTATCAAACAGGATTAGATACCCTGGTAGTCCA{2}CACGGTAAACGATGAATACT
CGCTGTTTGC{2}GATATACAGCAAGCGGCCAAGCGAAAGCGTTAAGTATTCCACCGTG
GGGAGTACGCCGGCAACGGTGAAACTCAAAGGAATTGGACGGGGGCCCGCACAAGCGGA
GGAACATGTGGT{3}TTAATTCGATGATACGCGAGGAACCTTACC{3}CGG

>Sample Read 2
ACCG{2}CACGGTAAACGATGAATACTCGCTGTTTGC{2}TGGCCCAGAACATCGCCTA
CCCACTGCAACTCGATCACTGGCGCAAGGACTATCAGGATCGTCGTGTCAACGAACTCCT
TGAATTCGTGGGCCTCAGTGAGCACGCAAACAAATACCCTTCGCAGCTGTCCGGCGGCCA
GAAG{1}AGCTGCAACTGACATTGAGGCTCGAAAGTG{1}CAGCGCGTCGGCATCGCCC
GCGCACTGGCCACTAATCCGGAGATTCTGCTCGCCGACGAAGCCAC{3}TTAATTCGATG
ATACGCGAGGAACCTTA{3}
```

Fig. 1. Two modified reads (Sample reads 1 and 2) are depicted as an example of problems that could arise from k-mer-based approaches used in software such as COMMET. Notice that the 3 colored k-mers (surrounded by brackets with IDs) are found in both sequences but in different places. In this example, COMMET would classify the sample read 1 as being equal to sample read 2 since it requires T non overlapping k-mers of length K to accept the equalness. The match is accepted even with varying parameters for $K \leq 30$ and $T \leq 3$, which includes default parameters. The sample read 1 has been extracted from the 16S reads contained in the sample run SRR029687 (https://www.ebi.ac.uk/metagenomics/projects/SRP000319/samples/SRS000998/runs/SRR029687/results/versions/1.0). Sample read 2 has been extracted from the full collection of reads contained in the same run.

4 Results and Discussion

In order to show the difficulty of assessing similarities between two metagenomes, a comparison was performed using two samples from the Human Microbiome Project database (HMP). In particular, the runs SRS014475 (in advance M1) and SRS015062 (M2), which correspond to reads extracted from the throat of healthy humans. The details of the sequencing machine used, filtering and trimming protocols can be found at the HMP website[2]. Both metagenomes were compared using COMMET and our proposed approach by computing the number of shared reads and the Jaccard Index. MASH and SIMKA were not included since they do not offer a read-level similarity measure. Additionally, a gapped BLASTn run was executed with a minimum expected value in the alignments of 10^{-5} to contrast results in terms of the number of shared reads and the Jaccard Index. COMMET was run using different parameters (the k-mer size ranging from 20 to 30 and the number of non overlapping k-mers t needed to accept a match ranging from 2 to 3). The proposed method was run using default parameters (5 for open gap penalty and 2 for extension, $+4$ and -4 for match and mismatch, respectively).

[2] http://hmpdacc.org/HMASM/#data.

Table 1 shows the number of reads reported to be included from the sample M1 in M2 using our proposed method, COMMET and BLASTn as a reference. Notice that COMMET was run with only one thread since it requires a SGE cluster to run in parallel, which unfortunately is a resource not available at our testing facilities. On the other hand, BLASTn and IMSAME were run using 30 threads since the computation of gapped alignments requires more execution time and their parallelization strategy does not need a specific cluster platform. Although COMMET is reporting a higher number of matched reads for the execution $k = 2$ and $t = 2$ compared to the rest of the applications, it is hard to assess whether these results represent the real similarity between the samples, since a slight change of parameters significantly changes the number of shared reads. On the other hand, our approach showed a higher number of matches than BLASTn with high percentage of identity and coverage (over 80%). Additionally, the gapped alignments were written to the output file to enable careful examination. The proposed method uses the percentage of identity and coverage as indicators of the alignment quality. The parameter values are on the researchers choice depending on the precision they wish to obtain. However, new quality indicators can be easily incorporated. Furthermore, the parameters on which COMMET is based do not allow to set up an experiment with ease, since the recommended size of k and t can strongly change depending on several factors such as the type of reads that are being compared, the type of metagenomes, the machine used to sequence the samples, etc. In the case of BLASTn, the default parameters do not facilitate the insertion of gaps in small sequences, thus producing almost ungapped results. In this sense, the developed method is strictly

Table 1. The number of matched reads is depicted depending on the program used and the set of parameters, along with the time in minutes and the Jaccard similarity index. COMMET is able to compute faster, but the results show high variation with a standard deviation of 188,239 in the number of reads compared to that of IMSAME, 31,326.

Program	Reads matched	Time (minutes)	Jaccard index
BLASTn -evalue 10^{-5}	428,366	469	0.28
BLASTn -evalue 10^{-5} -k 20	391,350	213	0.26
BLASTn -evalue 10^{-5} -k 30	364,889	243	0.24
IMSAME -p 80 -evalue 10^{-5}	546,157	45.5	0.36
IMSAME -p 90 -evalue 10^{-5}	513,079	56.4	0.34
IMSAME -p 95 -evalue 10^{-5}	483,537	158	0.32
COMMET -k 20 -t 2	626,788	45.89	0.41
COMMET -k 20 -t 3	335,513	46.84	0.22
COMMET -k 30 -t 2	311,718	0.58	0.20
COMMET -k 30 -t 3	177,353	1.2	0.11

intended for metagenome-metagenome comparison scenarios, and thus a flexible set of penalty costs should be used to model evolutionary events.

5 Conclusions

In this work we showed the strengths and weaknesses of current metagenome-metagenome comparison software. In addition, we performed an analysis of common approaches for metagenomic studies (BLAST and COMMET) at read level. We showed that COMMET is able to run very fast, but producing highly variable results whose validity is hard to assess in terms of the precise assignation of reads. In addition, we showed that there is no specific software for fine-grained reads to reads alignment. The developed method is able to compute gapped alignments between reads, enabling the modelling of the highly variable microbial communities present in metagenomes. Our approach is able to take advantage of the massively parallel architectures that are available nowadays, which enables our software to compute in reasonable times while maintaining a sensitive and robust detection of matches. We also present an incremental aligning method to reduce running times composed of alignment-free methods, gapped-free methods and finally gapped alignments. Additionally, the developed method is able to report the results at read level, reporting the alignments and thus enabling a careful examination. In terms of future work, we are aiming to produce a distance metric between metagenomes to approximate the number of species present based on clustering methods.

References

1. Maillet, N., et al.: Compareads: comparing huge metagenomic experiments. BMC Bioinform. **13**(19), 1 (2012)
2. Benoit, G., et al.: Multiple Comparative Metagenomics using Multiset k-mer Counting. arXiv preprint arXiv:1604.02412 (2016)
3. Vinga, S., Almeida, J.: Alignment-free sequence comparisona review. Bioinformatics **19**(4), 513–523 (2003)
4. Bonham-Carter, O., Steele, J., Bastola, D.: Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis. Brief. Bioinform. **15**(6), 890–905 (2013). doi:10.1093/bib/bbt052
5. Lippert, R.A., Huang, H., Waterman, M.S.: Distributional regimes for the number of k-word matches between two random sequences. Proc. Natl. Acad. Sci. **99**(22), 13980–13989 (2002)
6. Anne, C., et al.: A new statistical approach for assessing similarity of species composition with incidence and abundance data. Ecology Lett. **8**(2), 148–159 (2005)
7. Ondov, B.D., et al.: Mash: fast genome and metagenome distance estimation using MinHash. BioRxiv (2016). doi:10.1101/029827
8. Yuriy, F., et al.: How independent are the appearances of n-mers in different genomes? Bioinformatics **20**(15), 2421–2428 (2004)
9. Maillet, N., et al.: COMMET: comparing and combining multiple metagenomic datasets. In: 2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE (2014)

10. Langmead, B., Salzberg, S.L.: Fast gapped-read alignment with Bowtie 2. Nature Methods **9**(4), 357–359 (2012)
11. Altschul, S.F., et al.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**(17), 3389–3402 (1997)
12. Arjona-Medina, J.A., Torreno, O., Chelbat, N., Trelles, O.: Experimental Study of Local Alignment Distributions in the Comparison of Large Genomic Sequences Soibio (2013)