LAST + MEGAN-LR Approach to the Oxford Nanopore Wiggle Space Challenge

Caner Bagci[1,4] Benjamin Albrecht[1] Dominic Boceck[1] Ania Gorska[1,5]
Dino Jolic[4,5] Irina Bessarab[3] Rohan Williams[2,3] Daniel H. Huson[1,2,*]

1. Center for Bioinformatics, University of Tübingen, Sand 14, 72076 Tübingen, Germany
2. Life Sciences Institute, National University of Singapore, 28 Medical Drive, Singapore 117456
3. Singapore Centre for Environmental Life Sciences Engineering, National University of Singapore, 28 Medical Drive, Singapore 117456
4. Max-Planck Institute for Developmental Biology, 72076 Tübingen, Germany
5. IMPRS 'From Molecules to Organisms', Tübingen, Germany

Background

Metagenomics is the study of collection of microbial genomes from organisms populating a sample, using DNA sequencing [1]. The field has a broad range of applications, including medicine and environmental studies, and the aim is usually the taxonomic identification of microbes and/or their functional characterization [2, 3].

Short read NGS (Next Generation Sequencing) technologies, such as the Illumina HiSeq™ and MiSeq™ platforms, are widely used in metagenomic studies. However, rapid or even real-time and on-site identification of microbial communities and their functional characterization is of fundamental importance in some scenarios; for example, the identification of pathogens in a hospital setting [4].

Oxford Nanopore Technologies' MinION is a portable, realtime, single molecule sequencer that has been under development for the past five years. The reads produced by the MinION are "long reads", with an average length of 10 kb and a maximum length of close to one megabase (at present). These reads are much more error-prone than the short reads produced by NGS methods.

Another feature of the MinION is that individual reads become accessible as soon as they have been sequenced and so the downstream analysis can be performed while the device continues to sequence more reads. This feature will allow researchers to obtain answers to their questions almost in real time.

There exist many different computer programs and webservers for analyzing metagenomic datasets, aiming at taxonomic and/or functional binning and/or profiling the given data, such as MEGAN [5], MG-RAST [6] or Kraken [7]. Such tools are designed and engineered to work fast and accurately on short reads. Long reads pose a different set of challenges and require that existing approaches be adapted to address them. Here, we present a new variant of the lowest common ancestor (LCA) algorithm that is designed to perform taxonomic binning of long reads, and we demonstrate the use of the method on datasets provided by "The Oxford Nanopore 'Wiggle Space' Challenge" by CAMDA 2017, and on simulated Nanopore reads.

Methodology

Analysis Pipeline

We extracted reads in FastA format from base-called Fast5 files of the "wiggle space challenge" using poretools [8]. We have not yet made use of Illumina data associated with some of the MinION runs so as to perform

*To whom correspondence should be addressed.

error correction or hybrid assembly. We only consider microbial datasets that contain over 3000 reads, as the others do not produce a sufficient number of alignments.

For taxonomic binning and functional characterization of long-reads, we use a homology based LCA approach. We use LAST (v847) [9] to align all reads to the NCBI-nr (downloaded on 24.04.2017) protein reference database. Although LAST requires more than a day to build the index for NCBI-nr on a server, once the index is built it is extremely fast (aligning hundreds of long-reads per minute using 32 cores). In addition to its speed, it is also capable of producing frame-shift alignments, which is crucial for the analysis of MinION reads, given the current high rate of erroneous insertions and deletions in nanopore sequencing. For the challenge data, we ran LAST with the setting `-F 5`, setting the frameshift cost to 5, and otherwise using default settings.

MEGAN [5], our tool for taxonomic- and functional binning, and interactive exploration and visualization of metagenomic data, has been optimized to work efficiently on large numbers of large short read datasets. A number of the optimizations, and all binning algorithms, are based on the assumption that any given read is so short that overlaps at most one gene significantly. This assumption usually does not hold for long reads.

To address this, we have implemented a number of new algorithms and features that are specifically designed for long reads and that are bundled in a new long read (LR) mode of MEGAN, available since release 6.8.0 (May 2017).

As mentioned above, we propose to use the program LAST to compare all long reads against the NCBI-nr protein database. The output of LAST is then parsed and analyzed within MEGAN using the LR mode. We will refer to this analysis pipeline as LAST+MEGAN-LR.

For taxonomic binning of long reads, we have developed a binning algorithm called the *cover-based LCA*. Input is a long read r and the set M of alignments of the r to a protein reference database. We will describe the algorithm in two parts. First, for a given read r we will describe how to assign a weight for every taxon to which the read aligns. Then we will discuss how to place r on a taxonomic node using the weighted LCA algorithm.

Let $r = r_1 r_2 \dots r_n$ be a long read and let M be a set of alignments of r against a protein reference database. For any alignment $a \in M$, let $(a) = \text{bitscore}(a)/\text{length}(a)$ denote the bit score per base. Let M_t be the set of alignments that are associated with a given taxon t. For any position i in r, let $S(i) \subseteq M$ be the set of alignments the cover the position and are deemed significant with that property (E.g. we may require that the bit score of any alignment in $S(i)$ lies within 10% of the best score in $S(i)$. For any taxon t, let M_t be the set of all alignments in M that are associated with taxon t.)

For each position i in r and any taxon t associated with an alignment in M, we define the *position-taxon weight* as

$$\omega(i,t) = \frac{\sum_{a \in S(i) \cap M_t} h(a)}{\sum_{a \in S(i)} h(a)},$$

that is, the proportion of significant alignments that cover position i, and are associated with taxon t, weighted by bit score per base.

For each taxon t that is associated with any alignment in M, we define the *cover-weight* for r and t as

$$\omega(r,t) = \sum_{i=1}^{n} \omega_i(t),$$

that is, the sum of position-taxon weights for r and t.

The second part of the algorithms consists simply of placing the read r on the lowest taxonomic node that lies above a fixed percentage (50%, say) of the sum of all cover-weights computed for taxa to which the read aligns. (This is sometimes called the *weighted-LCA* algorithm.)

In summary, for a given read, the covered-base LCA determines weights for taxa to which the read aligns, based on amount of read that the alignments cover (weighted by bit scores per base), and then uses the weighted LCA to place the read.

Functional binning of short reads is usually performed by assigning each read to a class in a functional classification system such as InterPro [10] or eggNOG [11], based on its alignments.

This is often done using a simple *best-hit* strategy, as follows. For a short read r, let S be the set of significant alignments of r to a protein reference database such as NCBI-nr. Now, determine the highest-scoring alignment a in S for which the appropriate functional class is known and assign the read to that class. Note that any read is assigned to at most one functional class (and many reads are not assigned to any class, because the corresponding reference proteins are of unclassified.)

A long read may contain multiple genes and so we must modified the best-hit strategy to accommodate this. Let r be a long read and let M be the set of alignments computed for r. Detect the set of all segments $s_1, \ldots, s_m$ along the read where alignments pile up, ideally such that each segment corresponds to a different gene. For each such segment s_i, determine the set of significant alignments S_i that cover the segment. Here, significance is based bit score per base, to account for large differences in alignment length and an alignment must lie within 10% (say) of the best bit score per based on the given segment. Each segment s_i is then assigned to a functional class using the best-hit strategy applied to S_i.

In this study, we set the `PercentToCover` parameter in MEGAN to 50.0 and otherwise used the default parameters in long read mode.

Results and Discussion

We evaluated the performance of our approach on selected "wiggle space challenge" datasets, and on synthetic Nanopore datasets generated using nanosim [12]. For challenge datasets, for which the source organisms are known, we assumed that the true positive binning of reads is to exactly these organisms. We calculated sensitivity as the rate of true positives in total amount of data that is available; and specificity as the rate of true positives in total amount of data that was reported (TP / TP + FN). MEGAN can report either the number of reads assigned per taxa, or the sum of aligned base pairs for all the assigned reads. Because of the great variability of the length of long reads, we used the assigned base pair measure (Figure 1).

Our sensitivity on the challenge datasets is usually quite low. This is mainly due to the error-prone Nanopore reads not producing any alignments at all in most of these datasets. The Staphylococcus_aureus run, for example, has 74748 base-called reads, whereas LAST reported an alignment for only 4715 of them. We will look into improving our sensitivity at assigning error-prone long reads to taxonomic nodes by

- possibly improving base-calling using new base-callers, or
- by employing the provided Illumina datasets so as to error-correct the long reads.

Our analysis of the mystery data was severely hampered by the fact that 85% of the reads do not produce any alignments. Among those that aligned to reference sequences, a large portion was assigned to human and *Escherichia coli* (Figure 2). A significant amount of reads were also assigned to *Plasmodium ovale*, a parasitic protozoa. However, exploration of those reads, using MEGAN's new long-read inspector tool, casts doubts on their assignments as many of them also have alignments to *Staphylococcus aureus* and *Streptococcus pneumoniae*, both of which are also parasitic. We intend to improve our analysis of this dataset by extracting the reads assigned to these nodes and aligning them against genomes of these suspected parasites instead of the proteins. We also plan to compute a functional characterization of this dataset.

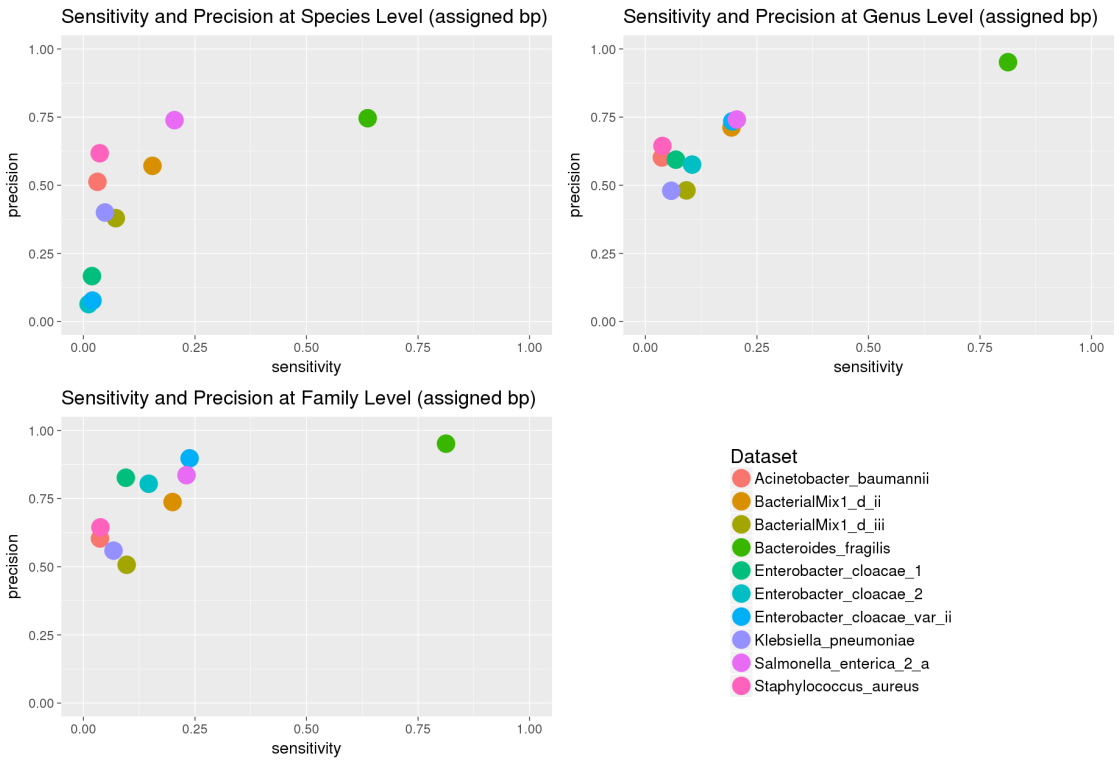

Figure 1: Sensitivity and Precision of LAST+MEGAN-LR computed for aligned base-pairs assigned at species, genus and family levels.

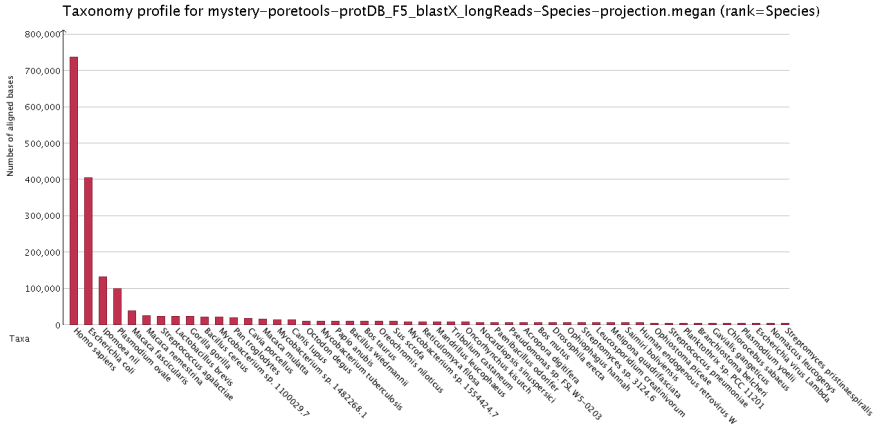

Figure 2: Bar chart of the top 50 taxonomic bins calculated for the mystery data identified by LAST + MEGAN-LR. Assignments have been projected to species level and only species are shown.

We simulated 2 hypothetical metagenomic datasets. The first is based on nine species of the family *Enterobacteriaceae*, from three different genera. We simulated 1000 reads from each species. Here, our aim was to create a very challenging situation for homology based methods as *Escherichia coli* has a vast amount of reference sequences in biological databases, which creates a huge bias towards it. LAST + MEGAN-LR was heavily affected by this problem and almost all reads from *Shigella* and many from *Cronobacter* genera either got assigned *Escherichia coli* or to higher taxonomic nodes.

The second simulation involved 2000 reads for five species randomly chosen from NCBI's Prokaryotic RefSeq Genomes database (Table 1). The performance of LAST + MEGAN-LR on this easy dataset was general very good, except for *Edwardsiella tarda*, which we could identify only at genus level. We would like to note that the remaining 25 to 30% of base-pairs that were not assigned are due to random sequencing artefacts in the Nanopore simulation, which is modelled in nanosim, rather than unreported false-negative alignments.

These datasets are available at `http://ab.inf.uni-tuebingen.de/data/software/megan6/long-read-data`.

Table 1: Percentage of assigned and correctly assigned base pairs for simulated Nanopore reads of 5 organisms randomly chosen from NCBI's Prokaryotic RefSeq Genomes.

			% correctly at level		
Species	**total bases**	**% assigned**	**species**	**genus**	**family**
Anoxybacillus-amylolyticus	15672248	78.3	50.9	68.5	76.9
Edwardsiella tarda	15600545	78.9	0.2	64.8	65.7
Methanobrevibacter-sp.-YE315	15336947	78.9	74.2	78.8	78.8
Methanocella paludicola	15678389	76.4	75.7	76.2	76.2
Myxococcus-hansupus	15608566	81.3	59.6	79.7	79.9

References

[1] P. Hugenholtz and G. W. Tyson, "Microbiology: metagenomics," *Nature*, vol. 455, no. 7212, pp. 481–483, 2008.

[2] W. R. Streit and R. A. Schmitz, "Metagenomics–the key to the uncultured microbes," *Current opinion in microbiology*, vol. 7, no. 5, pp. 492–498, 2004.

[3] T. M. Kuntz and J. A. Gilbert, "Introducing the microbiome into precision medicine," *Trends in Pharmacological Sciences*, vol. 38, no. 1, pp. 81–91, 2017.

[4] S. Juul, F. Izquierdo, A. Hurst, X. Dai, A. Wright, E. Kulesha, R. Pettett, and D. J. Turner, "What's in my pot? real-time species identification on the minion," *bioRxiv*, p. 030742, 2015.

[5] D. H. Huson, S. Beier, I. Flade, A. Górska, M. El-Hadidi, S. Mitra, H.-J. Ruscheweyh, and R. Tappu, "Megan community edition-interactive exploration and analysis of large-scale microbiome sequencing data," *PLoS Comput Biol*, vol. 12, no. 6, p. e1004957, 2016.

[6] E. M. Glass, J. Wilkening, A. Wilke, D. Antonopoulos, and F. Meyer, "Using the metagenomics rast server (mg-rast) for analyzing shotgun metagenomes," *Cold Spring Harb Protoc*, vol. 2010, pp. pdb.prot5368+, Jan. 2010.

[7] D. E. Wood and S. L. Salzberg, "Kraken: ultrafast metagenomic sequence classification using exact alignments," *Genome biology*, vol. 15, no. 3, p. R46, 2014.

[8] N. J. Loman and A. R. Quinlan, "Poretools: a toolkit for analyzing nanopore sequence data," *Bioinformatics*, vol. 30, no. 23, pp. 3399–3401, 2014.

[9] S. M. Kiełbasa, R. Wan, K. Sato, P. Horton, and M. C. Frith, "Adaptive seeds tame genomic sequence comparison," *Genome research*, vol. 21, no. 3, pp. 487–493, 2011.

[10] A. Mitchell, H.-Y. Chang, L. Daugherty, M. Fraser, S. Hunter, R. Lopez, C. McAnulla, C. McMenamin, G. Nuka, S. Pesseat, A. Sangrador-Vegas, M. Scheremetjew, C. Rato, S.-Y. Yong, A. Bateman, M. Punta, T. K. Attwood, C. J. Sigrist, N. Redaschi, C. Rivoire, I. Xenarios, D. Kahn, D. Guyot, P. Bork, I. Letunic, J. Gough, M. Oates, D. Haft, H. Huang, D. A. Natale, C. H. Wu, C. Orengo, I. Sillitoe, H. Mi, P. D. Thomas, and R. D. Finn, "The InterPro protein families database: the classification resource after 15 years," *Nucleic Acids Research*, vol. 43, no. D1, pp. D213–D221, 2015.

[11] S. Powell, D. Szklarczyk, K. Trachana, A. Roth, M. Kuhn, J. Muller, R. Arnold, T. Rattei, I. Letunic, T. Doerks, L. J. Jensen, C. von Mering, and P. Bork, "eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges," *Nucleic Acids Research*, vol. 40, no. Database-Issue, pp. 284–289, 2012.

[12] C. Yang, J. Chu, R. L. Warren, and I. Birol, "Nanosim: nanopore sequence read simulator based on statistical characterization," *bioRxiv*, p. 044545, 2016.