MCUIUC – A New Framework for Metagenomic Read Compression

Jonathan G. Ligo*, Minji Kim*, Amin Emad, Olgica Milenkovic and Venugopal V. Veeravalli

*The first two authors contributed equally to this paper.
Department of Electrical and Computer Engineering, University of Illinois, Urbana-Champaign
Email: {ligo2,mkim158,emad2,milenkov,vvv} [at] illinois.edu

Abstract—Metagenomics is an emerging field of molecular biology concerned with analyzing the genomes of environmental samples comprising many different diverse organisms. Given the nature of metagenomic data, one usually has to sequence the genomic material of all organisms in a batch, leading to a mix of reads coming from different DNA sequences. In deep high-throughput sequencing experiments, the volume of the raw reads is extremely high, frequently exceeding 600 Gb. With an ever increasing demand for storing such reads for future studies, the issue of efficient metagenomic compression becomes of paramount importance. We present the first known approach to metagenomic read compression, termed MCUIUC (Metagenomic Compression at UIUC). The gist of the proposed algorithm is to perform classification of reads based on unique organism identifiers, followed by reference-based alignment of reads for individually identified organisms, and metagenomic assembly of unclassified reads. Once assembly and classification are completed, lossless reference based compression is performed via positional encoding. We evaluate the performance of the algorithm on moderate sized synthetic metagenomic samples involving 15 randomly selected organisms and describe future directions for improving the proposed compression method.

I. INTRODUCTION

Metagenomics is an area of microbial genomics devoted to the unified study of complex cultures found in the human body and the ecosystem. The main difference between classical genomics and metagenomics is the fact that in the latter, no attempt is made to separate the organisms in a sample. The reasons for using such a holistic approach are either of purely technical nature – in cases when single sample isolation is technically implausible to accomplish – or they may be governed by the desire to explore the interactions, dynamical co-regulation and joint evolution of microbes in a given environment [14].

Unlike the field of genomics, metagenomics is still in its infancy, despite recent intense research efforts in the areas of metagenomic sampling, sequencing and assembly [14], [13], [11]. The main challenges in the field are a still fairly large cost of high-coverage metagenomic sequencing, although cost constraints are likely to diminish in the near future; and the extremely large file sizes produced by High-Throughput Sequencers such as Illumina, when applied to metagenomic data. As an illustration, the size of the human gut microbial population is estimated to be several hundreds, while average bacterial genome lengths are between 100 Kbp (Kilobasepairs) and 12 Mbp (Megabasepairs). With a coverage of at least 100 reads per base pair, one arrives at a rough estimate of 100 Gbp (Gigabasepairs) of genomic data. Analyzing, storing and transmitting such volumes of data is a formidable task, and most efforts on metagenome assembly require powerful and parallelized computer clusters to perform even the most basic operations.

In parallel, an extensive effort is underway to develop efficient means for lossless and lossy compression of whole genomes [16], as well as for compression of reads sampled from a single genome [8]. Specialized methods for compressing protein-coding regions as well as regions with frequent repeats were reported in [11], [3], [9], [13], using methods as diverse as modified Lempel-Ziv encoding, the Burrows-Wheeler transform and wavelet-based decomposition. Compression of reads is mainly achieved by using a reference genome, with an initial study based on Golomb codes reported in [8] and extensions thereof reported in [10].

We propose the first algorithmic solution to the problem of de novo metagenomic compression that mitigates the need for computationally demanding full metagenome assembly by using a read classification technique and subsequent reference-based single genome compression. The gist of the classification method is to use a microbe identification tool – Metaphyler [7] – that can accurately identify the order of a high percentage of organisms from a metagenomic mix. By aligning the reads to all genomes of organisms within the identified orders via Bowtie2 [5], one can provide a rough classification map for reference-based compression. For moderate size samples, unaligned reads usually amount to less than 20–25% of the reads, and can be assembled through existing metagenome assembly software such as IDBA-UD [11]. Starting positions of aligned reads and their differences from the reference genome are encoded using techniques outlined in [8], [19].

The paper is organized as follows. In Section II we describe the first algorithmic solution for metagenomic read compression. In Section III we demonstrate the performance of the method on synthetic Illumina sequencer data, using a randomly selected set of 15 bacterial organisms. Concluding remarks are given in Section IV.

II. ALGORITHMIC SOLUTION FOR METAGENOMIC COMPRESSION

Throughout the paper, we refer to an ordered sequence of symbols from the alphabet \( \{A, T, G, C\} \) as a genome or genomic sequence. Elements of the sequence are indexed from 1 to \( L \), the length of the sequence. A read is a substring of a genome, generated by some sequencing system. The starting and terminal location of a read correspond to the index of the
sequenced no matter how large the coverage [15].

Including the error rate of the sequencer, the structural properties of the assembly refers to the process of overlapping reads – suffix to the sequence are depicted in Figure 1. A similar alignment involving two genomes is depicted in Figure 2, also depicting several misaligned reads between species.

Given that metagenomic data carries valuable information about the population dynamics of microbes in a given environment, even small aberrations in the composition of the sample have to be observed and recorded for in-depth study of the underlying phenomena. Also, given the cost of acquiring samples to sequence and laboratory work necessary for sequencing, it is also valuable to preserve and share the sequencing experiment in its entirety. Hence, the goal of any compression method should be to preserve the reads intact, i.e., to perform lossless read compression. The main issues associated with individual lossless read compression is that the reads – as generated by Illumina, Roche454, IonTorrent – are usually of length not exceeding $L = 400 - 500$ bps, and that the number of reads is extremely large. The former property of reads does not allow any known compression algorithm to take full advantage of inherent patterns or data structures, since such algorithms offer good performance only for sufficiently long sequence lengths [4]. On the other hand, the existence of a huge volume of reads creates the problem of efficiently using their overlaps. Such overlaps arise due to the requirement of large genome coverage, given that large coverage is needed for accurate sequence reconstruction[7].

One simple method to mitigate the aforementioned problems is to perform reference based compression [8]. Reference based compression is based on the assumption that one knows some rough form of the genome the reads were generated from, without knowing the exact identity of the organism. Such an approach cannot be used for de novo compression, where the origin of the reads is unknown. Such is the case for most metagenomic datasets currently available.

One obvious solution would be to first apply a metagenomic assembly process to create an adequate number of reference genomes, and subsequently compress all the reads using the “closest” genome as a reference. This is infeasible in the long term due to the reduction of costs in sequencing technology, increases in sequencing depth and a desire to sequence more complex metagenomic samples. A current Next Generation sequencer such as the Illumina HiSeq 2500 system can produce up to 4.5 Gb/hr [21], a rate which will be quickly exceeded by future sequencers. Thus, a laboratory running multiple sequencers must archive and distribute petabytes of data per year. Since the read lengths are short, techniques such as de Brujin graph assemblers [11] are needed for assembly at the cost of high memory and CPU time for large data sets, requiring prohibitively expensive computers. One approach to mitigate a related problem via distributed computing is described in TIGER [17] for large single genomes, though this approach has not been validated on metagenomic samples. The insight towards the composition of the metagenomic sample is also deferred until after assembly. One approach that mitigates this issue – and the one pursued in this paper – is to perform a rough classification of the reads into subgroups that may individually be processed more efficiently for the purpose of both identifying a proper reference and for subsequent compression. The structure of such an algorithm is depicted in Figure 5 with a brief description of the blocks provided in what follows.

- **Step 1 (Classification):** The first step of the procedure is a “rough” identification of the mixture of species in the metagenomic sample. One well established method for bacterial genome identification is the use of the so called 16S rRNA regions [6]. Although this method is helpful for basic taxonomic identification, it does not allow for the precision of methods based on multiple marker identification, such as MetaPhyler [7]. The gist of the approach in [7] is that almost every genomic substring of length exceeding 20 is unique to a species. MetaPhyler hence scans for markers in excess of this length and links them to markers of documented species. Therefore, known organisms have a high chance of being identified by this method up to a given taxonomic order, in this case, their genus. As will be shown in the next section, MetaPhyler reports the number of identified markers of genera as well as their corresponding abundance levels, but the algorithm often misses a number of species (most often, undocumented species) while introducing a large number of false positives. In order to optimize the performance of the subsequent steps of the algorithm, the selection of genera most likely contained in the mixture has to be performed using the identified number of markers, the length of the genome and other criteria. We elaborate on this issue in the following section.

- **Step 2 (Partitioning of the Dataset):**
  1. Once a group of genera is selected according to the Procedure in Step 1, a set of reference genomes “best” aligned to the reads has to be selected. Given that MetaPhyler produces predictions only up to the level of genera, one possi-
ble way to identify representative genomes in the mixture is to select all species of the identified genera and perform quick testing of alignment quality with the given reads. This task can be accomplished by using Bowtie2 [5], a specialized software for ultra-fast alignment of short reads to long genomes. Bowtie2 was designed as an alternative to the BLAST (Basic Local Alignment Search Tool) [2] for cases when one has a very large number of short query sequences to be aligned with a small number of genomes. Running Bowtie2 on the metagenomic reads and reference genomes from all species within the identified genera allows for performing read classification up to the genus level in terms of best alignment score. It is worth pointing out that some reads will be reported as unaligned, i.e., as not having sufficiently high similarity to any of the species to which alignment was performed. A means for parallelizing single genome assembly that shares some of the classification ideas of Step 2 was first reported in [17].

2) In order to keep the number of reference genomes used for compression reasonably small, the next step is to combine the reads assigned to members within one single genus, independent on which species they were identified to belong to. The species of each genus which had the most reads aligned to it is called the representative of the genus and is used for reference-based compression of all the reads mapped to the genus.

3) Unaligned reads are treated differently. Given that their number is relatively small compared to the number of aligned reads, metagenomic assembly of such reads may be performed in a computationally efficient manner using a metagenomic assembler such as IDBA-UD [11]. The assembler produces contigs – overlapped strings of reads of long lengths – that may subsequently be queried in BLAST to identify the organisms they most likely originated from. Steps 2.1 and 2.2 are repeated using as input the unaligned reads from step 2.1 as well as the genera of organisms associated with the longest contigs.

- **Step 3 (Compression and Distribution):** The number of unaligned reads remaining after step 2.3 is relatively small and can be compressed using a standard tool such as bzip2. The reads associated with each genus are aligned to their representative using Bowtie2 and outputted into the SAM (Sequence Alignment/Map) format. The resulting SAM files are converted to the sorted and indexed BAM format (a binary format for sequencing data) via SAMTools [20]. The BAM file for each genus is compressed via reference-based compression against its representative to the CRAM format using the CRAM toolkit [19]. The CRAM toolkit provides a practical reference-based compression procedure and file format based on [8]. Note a simple extension to the algorithm akin to [8] is to use some of the contigs as references for compression. These references would be packaged for distribution. The compressed unaligned reads, CRAM files, list of representatives for the genera, the corresponding genomes (if they are not available through a standard database such as NCBI’s microbial genome database) and Metaphyler taxonomic classification are packaged into a tar archive. The resulting archive can be distributed and the reads losslessly reconstructed via the CRAM toolkit, given the CRAM files and representative genomes. The representative genomes may be compressed using standard compressors such as bzip2 or specialized compressors such as DNACompress [18].

### III. Working Example

Metagenomic samples come in various sizes – most environmental samples contain anywhere between tens to hundreds of species and subspecies. The human gut is estimated to contain roughly 300 bacteria, while human reproductive systems host roughly 20 – 40 microbes [13]. In order to illustrate our findings in a compact form, we perform a step-by-step analysis of the compression scheme described in the previous section on simulated metagenomic data containing 15 species.

#### A. Simulating the Metagenomic Sample

Species were randomly selected from the NCBI microbial genome database available at [22]. The se-
TABLE I. A RANDOMLY SELECTED SET OF 15 SPECIES USED TO ILLUSTRATE THE OPERATING PRINCIPLES OF MCUIUC.

| # | Genus                 | Species                                     |
|---|-----------------------|---------------------------------------------|
| 1 | Aquifex               | Aquifex-aequorum-VF5_uid57765               |
| 2 | Bacillus              | Bacillus-molycoccus-fk-17352b               |
| 3 | Cycloclasticus        | Cycloclasticus-P1_uid176308                |
| 4 | Enterococcus          | Enterococcus-acidofaciens-D32_uid171261    |
| 5 | Escherichia           | Escherichia-coli-858_uid62979               |
| 6 | Francisella           | Francisella-molycoccus-U12_uid58499         |
| 7 | Geobacter             | Geobacter-P71_uid16533                     |
| 8 | Helicobacter          | Helicobacter-P71_uid16533                  |
| 9 | Lactobacillus         | Lactobacillus-acidophilus-30SC_uid56065     |
| 10 | Lacticoccus           | Lacticoccus-lactis-CV56_uid160253          |
| 11 | Mycoplasma            | Mycoplasma-genitalum-G37_uid37707          |
| 12 | Polynucleobacter      | Polynucleobacter-necessarius-SHR1_uid15587 |
| 13 | Oscillatoria          | Oscillatoria_PCC_7112_uid183110            |
| 14 | Thermococcus          | Thermococcus-4557_uid70841                 |
| 15 | Vibrio                | Vibrio-splendidus-15GB32_uid9353           |

Metaphyler takes the simulated metagenomic reads as inputs and outputs their taxonomy classifications. We focused our attention on genus classification, as it is the most specific level provided by Metaphyler. The Metaphyler genus-level output for the input metagenomic reads is given in Fig. 4. Of the 31 genera identified, we selected only those with percent abundance higher than a threshold, which in this case was chosen as 0.1. Note that the choice of the threshold is governed by many parameters, including the number of estimated organisms, their genome length, the number of known markers in the genomes, as well as the actual output of Metaphyler. A detailed analysis of the threshold selection method is relegated to the full version of the paper. The selected set of 14 genera contains 13 true positives and one false positive (note that Synechococcus was not part of the selected 15 species). In addition, Metaphyler missed identifying two organisms present in the metagenomic sample: Cycloclasticus_P1_uid176368 and Oscillatoria_PCC_7112_uid183110.

Given that Metaphyler produces only genus-level classifications, it remains to identify the species within the metagenomic mixture. We tried two approaches for addressing this problem: in the first method, we randomly selected a small number of species (2-8, or the largest number of documented species if this number was smaller than eight) for each of the identified genera. In the second approach, we selected all known species within each chosen genera. As an illustration, more than 35 strains of E. coli are sequenced so far, with more than 20 other sequencing efforts close to completion, while only one species of Cycloclasticus has been sequenced so far (see Wikipidea – ecoliwiki – for more information). In both cases, the genomes of the selected species were used as references for classification, as described in the next subsection.

C. Read Classification

For the purpose of classifying the reads based on similarity to the selected reference genomes, we used the Bowtie2 algorithm [5]. Bowtie2 is based on two concepts commonly encountered in computer science and information theory: the Burrows-Wheeler (BW) transform and suffix-tree searching. The BW transform is performed to index the reference genomes for subsequent read alignment. Searching for the best alignments is accomplished via suffix-trees, which makes the solution computationally efficient even for very large number of sequence reads.

The classification results were of poor quality when only 2-8 reference genomes from each genera were selected randomly, and are not reported. On the other hand, using all species of the 14 chosen genera above and building a Bowtie2 index provided very good metagenomic read alignment. In the first round of alignment, 78% of reads were grouped according to the species they were identified to belong to. In the second round of alignment, only one species per genera with largest number of aligned reads was used as a reference for the genera. The unaligned reads from the first round were assembled using IDBA-UD. The resulting contigs were passed through BLAST. As an illustration, BLASTing the 30 longest contigs (of lengths 31k to 722k) led to identification of the two species missed by Metaphyler with very high confidence level: 12 of the contigs came from Cycloclasticus_NC018697, while 18 contigs were identified as coming from Oscillatoria_NC019729. Subsequently, unaligned read classification was performed for the two newly identified species, following the same steps as outlined in connection with the Metaphyler output processing scheme.

D. Compression

For comparison, the sample was initially compressed with gzip and bzip2 resulting in 1.4 GiB and 1.3 GiB file sizes,
As the number of reads mapped to Synechococcus is very small given the relative sizes of the detected genomes, it is reasonable to assume it was a false positive and lump the Synechococcus reads with the unaligned set for a clearer picture of the contents of the metagenomic sample. In addition, note that given the reference genome lengths, the proportion of each organism in the metagenomic sample can be estimated assuming uniform coverage.

**TABLE II. SIZE OF READS MAPPED TO EACH GENERA (UNIT: NEAREST MiB)**

| Identified via Metaphyler | FASTA | BAM | CRAM |
|--------------------------|-------|-----|------|
| Aquiflex                  | 181   | 45  | 15   |
| Bacillus                  | 458   | 115 | 40   |
| Enterococcus              | 349   | 87  | 30   |
| Escherichia               | 589   | 149 | 52   |
| Francissella              | 223   | 56  | 19   |
| Geobacter                 | 504   | 126 | 44   |
| Helicobacter              | 195   | 49  | 17   |
| Lactobacillus             | 243   | 60  | 21   |
| Lactococcus               | 280   | 70  | 24   |
| Mycoplasma                | 67    | 16  | 6    |
| Polymicrobial             | 182   | 45  | 16   |
| Synechococcus             | 1     | 0   | 0    |
| Thermococcus              | 235   | 59  | 20   |
| Vibrio                    | 386   | 96  | 33   |
| Total                     | 5046  | 1261| 436  |

**IV. CONCLUSION**

We described the first *de novo* metagenomic read compression algorithm. The algorithm was based on four major subroutines: genera identification, species identification, read classification and reference based single-genome compression. The performance of the method was illustrated on a simulated metagenomic sample with 15 species, for which overall compression ratio on the order of 12.5 were reported. Further work includes integrating side information such as genera known to be present in a sample with high likelihood, selection rules for using multiple representatives for a genus and integration of phylogenetic aligners to mitigate the “Other” category used in Metaphyler at the genus level.

**ACKNOWLEDGMENT**

This work was supported in part by NSF grants CCF 0809895, CCF 1218764, Emerging Frontiers for Science of Information Center, CCF 0939370 and U.S. Defense Threat Reduction Agency through subcontract 147755 at the University of Illinois from prime award HDTRA1-10-1-0086. The authors also gratefully acknowledge many useful discussions with Prof. Jian Ma and Xiaolong Wu at the University of Illinois, Urbana-Champaign.

**REFERENCES**

[1] D. Adjeroh et al., *DNA Sequence Compression Using the Burrows-Wheeler Transform*, Proceedings of the IEEE Computer Society Conference on Bioinformatics, 2002.

[2] L. Altschul et al. *Basic local alignment search tool*, Journal of Molecular Biology, 1990.

[3] S.K. Chilappagari, O. Milenkovic and B. Vasic, *DNA sequence analysis using fractal grammars and integer wavelet transforms*, Proc. IEEE Int. Workshop on Genomic Signal Process. and Statistics, 2005.

[4] T. Cover and J. A. Thomas, *Elements of information theory*, John Wiley Sons, 2012.

[5] B. Langmead and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2*, *Nature Methods*, 2012.

[6] W-T. Liu et al., *Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA*, *Applied and environmental microbiology*, 65, no. 11, 4516-4522, 1997.

[7] B. Liu et al., *MetaPhyler: Taxonomic profiling for metagenomic sequences*, *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2010.

[8] F. Markus et al., *Efficient storage of high throughput DNA sequencing data using reference-based compression*, *Genome research* 21, no. 5, pp. 734-740, 2011.

[9] C. G. Nevill-Manning and I. H. Witten, *Protein Is Incompressible*, Proceedings of the Conference on Data Compression, 1999.

[10] D. S. Pavlichin, T. Weissman, and G. Yona, “The human genome contracts again,” *Bioinformatics*, online, July 2013.

[11] Y. Peng et al. *IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth*, Nat Methods, 2012.

[12] P.A. Pevzner, H. Tang, and M.S. Waterman, *An Eulerian path approach to DNA fragment assembly*, *Proc Natl Acad Sci* USA, 2001.

[13] J. Qin et al., *"A human gut microbial gene catalogue established by metagenomic sequencing.* Nature 464, no. 7285, pp. 59-65, 2010.

[14] C. Riesenfeld, P. D. Schloss, and J. Handelsman, *Metagenomics: genomic analysis of microbial communities,* *Annu. Rev. Genet.*, 38, pp. 525-552, 2004.

[15] M-F Sagot, *Spelling approximate repeated or common motifs using a suffix tree,* In *LATIN'98: Theoretical Informatics*, pp. 374-390. Springer Berlin Heidelberg, 1998.

[16] N. Siva, “1000 Genomes project,” *Nature Biotechnology*, 26, no. 3, pp. 256-256, 2008.

[17] X-L. Wu et al., “TIGER: tiled iterative genome assembler,” *BMC bioinformatics* 13, no. Suppl 19, S18, 2012.

[18] C. Xin, M. Li, B. Ma, and J. Tromp, “DNACompress: fast and effective DNA sequence compression,” *Bioinformatics* 18, no. 12, pp. 1696-1698, 2002.

[19] European Molecular Biology Laboratory - European Bioinformatics Institute, “CRAM format Specification”, ver. 2.0.1. June 12, 2013 [Online]. Available: http://www.ebi.ac.uk/ena/sites/ebi.ac.uk.ena/files/documents/cram_format_2.0.pdf

[20] H. Li et al. and 1000 Genome Project Data Processing Subgroup, “The Sequence alignment/map (SAM) format and SAMtools,” Bioinformatics, vol. 25, pp. 2078-9, 2009.

[21] Illumina, Inc. (2013) “Performance and Specifications for HiSeq 2500/1500” [Online]. Available http://www.illumina.com/systems/hiseq_2500_1500_performance_specifications.html

[22] ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/all.fna.tar.gz