De novo likelihood-based measures for comparing metagenomic assemblies

Christopher M. Hill*†, Irina Astrovskaya*, Howard Huang‡, Sergey Koren§, Atif Memon†,
Todd J. Treangen§ and Mihai Pop*†
*Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, USA
Email: {cmhill,irina,mpop}@umiacs.umd.edu
†Department of Computer Science, University of Maryland, College Park, MD 20742, USA
Email: atif@cs.umd.edu
‡Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland, USA
Email: hhuang58@jhu.edu
§National Biodefense Analysis and Countermeasures Center, Battelle National Biodefense Institute, Frederick, Maryland, USA
Email: {sergek,treangen}@umiacs.umd.edu

***Abstract*—Metagenomic assemblers inherit all the difficulties of traditional single genome assembly, but with the additional complexity of trying to resolve assemblies of closely related species with drastically varying abundances. Assessing and comparing the quality of single genome assembly still relies on the availability of independently determined standards, such as manually curated genomic sequences. These standards are often not possible in metagenomic projects, where a large portion of the organisms and strains are novel. Thus, we must rely on *de novo* methods for assessing and comparing assembly qualities. Here we describe an extension to our *de novo* LAP framework that takes into account abundances of assembled sequences to accurately and efficiently evaluate metagenomic assemblies. We have integrated our LAP framework into the metagenomic analysis pipeline MetAMOS, allowing any user to reproduce quality assembly evaluations with relative ease.**

***Keywords*-metagenomics; *de novo* assembly evaulation;**

I. Introduction

The decreasing costs of sequencing technology has led to a sharp increase in metagenomics projects over the past decade. These projects allow us to better understand the diversity and function of microbial communities found in the environment, including the ocean [1], and the human body [2], [3]. Traditional *de novo* genome assemblers have trouble assembling these datasets due to the presence of closely related species and and the need to distinguish between true polymorphisms and errors arising from the sequencing technology. Metagenomic assemblers often use heuristics based on sequencing (Meta-IDBA [4] and MetaVelvet [5]) and k-mer (Ray Meta [6]) coverage to split the assembly graph into subcomponents that represent different organisms, then apply traditional assembly algorithms on the individual organisms.

As the number of metagenomic assemblers available to researchers continues to increase, the development of approaches for validating and comparing the output of these tools is of critical importance. Despite the incremental improvements in performance, none of the assembler tools available today outperforms the rest in all cases (as highlighted by recent assembly bake-offs GAGE [7] and Assemblathons 1 [8] and 2 [9]). Different assemblers "win" depending on the specific downstream analyses, structure of the genome, and sequencing technology used. Furthermore, evaluating the trade-off between increased contiguity and errors is difficult even when there is a gold standard reference genome to compare to, which is not available in most practical assembly cases. Thus, we are forced to heavily rely on *de novo* approaches based on sequence data alone.

One objective *de novo* metric that has been used to evaluate and compare assembly quality is based on the likelihood that the observed reads are generated from the given assembly, which can be accurately estimated by modeling the sequencing process. This metric was proposed by Gene Myers in his pioneering work in the 1990's, where he suggested that the correct assembly must be consistent with the statistical properties of the data generation process. This idea was extended and used by recent assembly evaluation frameworks: ALE [10], CGAL [11], and LAP [12].

Most of the previous *de novo* and reference-based validation methods have been designed for single genome assembly. Currently, there are no universally-accepted reference-based metrics for evaluating metagenomic assemblies. Despite reference sequences being available for a small fraction of organisms found in metagenomic environments [1], it is not clear how to distinguish errors from genomic variants found within a population. Furthermore, it is not clear how to weigh errors occurring in more abundant organisms. Likelihood-based frameworks, such as ALE [10], CGAL [11], and LAP [12], rely on the assumption that the sequencing process is approximately uniform across the genome; however, the sequencing depth across genomes in metagenomic samples can vary greatly [13].

In our paper, we describe an extension to our LAP framework to evaluate metagenomic assemblies. We will show that by modifying our likelihood calculation to take

978-1-4799-1310-7/13/$31.00 ©2013 IEEE

into account contig abundances, we can accurately and efficiently evaluate metagenomic assemblies. We evaluate our extended framework on data from the Human Microbiome Project (HMP). Finally, we show how our LAP framework can be used automatically by the metagenomic assembly pipeline MetAMOS [14] to select the best assembler for a specific dataset, and to provide users with a measure of assembly quality. The software implementing our approach is made available, open-source and free of charge, at: http://assembly-eval.sourceforge.net/ and with the MetAMOS package: https://github.com/treangen/MetAMOS.

II. Methods

A. Likelihood of an assembly

Our LAP framework measures the quality of an assembly as the probability that the observed reads, R, are generated from the given assembly, A: $\Pr[R|A]$] [12]. Assuming that the event of observing each read is independent, then the probability $\Pr[R|A]$ of the read set R being produced from the assembly A, is the product of the individual read probabilities, p_r. That is,

$$\Pr[R|A] = \prod_{r \in R} p_r \tag{1}$$

By modeling the data generation process, we can calculate the probability of each read, p_r. Assuming uniform coverage, where each position in the genome is covered by roughly the same amount of reads as any other position, a read may be sequenced starting from any position of the genome with equal probability. In the basic error-free model, if a read matches to one position in the assembly, and its reverse-complement does not match anywhere, then the probability of the read being produced from the assembly is $p_r = \frac{1}{2L}$, where L is the length of the assembly. The length of the assembly is doubled due to the double-stranded molecules of DNA that make up the genome. Thus, if a read matches at n_r positions on the assembly, then $p_r = \frac{n_r}{2L}$.

Ghodsi et al. details how to modify the calculation of p_r to handle practical constraints, e.g., sequencing errors and mate pairs. They also show that the true genome maximizes $\Pr[R|A]$ (see [12] for more details).

Calculating $\Pr[R|A]$ can be expensive for dataset sizes commonly encountered in sequencing projects (tens to hundreds of millions of reads). Thus, we can approximate the likelihood of the assembly by using a random subset of the reads (R'). To counteract the effect of sample size on the probability, we define the assembly quality (LAP score) as the log of the geometric mean of the read probabilities.

The mean of the read probabilities over the sample is expected to be equal to the mean over all reads, but if the sample size is too small, then the accuracy of the estimation will be poor.

B. Extending LAP to metagenomic assemblies

An important simplifying assumption of our framework is that the sequencing process is uniform in coverage. In metagenomics, however, the relative abundances of organisms are rarely uniform [13], reflecting the difference in abundance between the different organisms within a community. Here we show that taking this abundance information into account allows us to extend the LAP framework to metagenomic data. We now assume that while the abundances of each organism may vary dramatically, the sequencing process still has uniform coverage across the *entire* community. For example, consider a simple community containing two organisms (A and B), one which is twice as abundant as the other. This community, thus, comprises twice as much of A's DNA than that of B. Assume, for simplicity, that the community contains exactly three chromosomes (two of A and one of B). A random sequencing process would sample each of these equally, and an ideal metagenomic assembler would produce two contigs, one covered twice as deep as the other.

In essence, we view the collection of individual genomes and their relative abundances as a single *metagenome* where each genome is duplicated based on their abundance. This setting is similar to that of repeats in single genome assembly, where a repetitive element can now include an entire genome. Like in the case of single genomes, the assembler that correctly estimates these repeat counts maximizes the LAP score. In other words, in order to accurately evaluate the metagenomic assemblies using our LAP framework, the abundance (or copy number) of each contig is needed. As most metagenomic assemblers do not report this information, here we use the average coverage of the contig (provided by the MetAMOS pipeline) to represent the copy number. The median coverage of a contig can also be used to provide a more robust estimate of contig abundance.

In the error-free model, we compute the probability of a read, p_r, given the assembled sequence and abundance as:

$$p_r = \frac{\sum_{c \in \text{Contigs}} \text{abun}(c) * n_{rc}}{2\hat{L}} \tag{2}$$

where abun(c) is the abundance of contig c, n_{rc} is the number of times read r occurs in contig c, and $\hat{L}$ is the sum of contig lengths weighted by their abundance. In the case where the abundance of each contig is 1, calculating p_r is identical to the original LAP (single genome) formulation. A similar modification can be done to handle sequencing errors outlined in [12].

Our prior work has shown we can approximate the probabilities using fast and memory efficient search alignment programs (e.g., Bowtie2 [15]) when it is impractical to calculate the exact probabilities for large read sets. We can apply the metagenomics modification above to the alignment tool-based method:

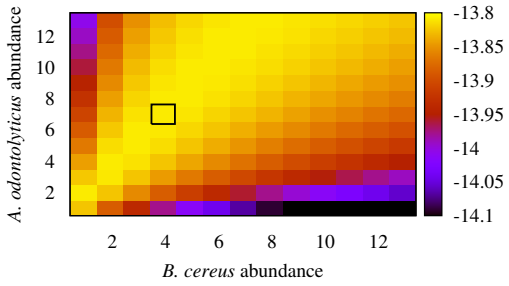

Figure 1. LAP scores for a simulated metagenomic community: *B. cereus* (4 copies, 5.2MB) and *A. odontolyticus* (7 copies, 2.4MB). Each cell (x,y) represents the LAP score for a mixture of *x* copies of the *B. cereus* and *y* copies of the *A. odontolyticus*. The true abundance ratio maximizes the LAP score (indicated by a black rectangle).

$$p_r = \frac{\sum_{j \in S_r} \text{abun}(j_{\text{contig}}) * p_{r,j_{\text{subs}}}}{2\hat{L}} \quad (3)$$

where S_r is the set of alignments in the SAM file for the read r and the probability of alignment, $p_{r,j_{\text{subs}}}$, is approximated by $\epsilon^{subs}(1-\epsilon)^{l-subs}$ where ϵ is the probability of an error (a mismatch, an insertion, or a deletion).

An important factor in any likelihood-based assembly evaluation framework is the handling of reads that do not align well to the given assembly. In practice, unalignable reads are often the result of sequencing errors and contaminants. If these reads are given a probability close to 0, than the best assembler would be the one that incorporates the most reads. To account for this, in our original LAP framework, any read that does not align well, the overall assembly probability does not decrease more than the probability of an assembly that contains the appended read as an independent contig (see [12] for more details). This does not change when we handle metagenomic data, since the average coverage of the "new" contig is one.

C. Integration into MetAMOS

In addition to being a standalone framework, the software implementing our metagenomic LAP approach comes packaged with the MetAMOS pipeline. We modified MetAMOS so users can now specify multiple assemblers (comma-separated) after the `-a` parameter, and `runPipeline` will run all assemblers and select the assembly yielding the highest LAP score to be used in downstream analyses.

III. Results

A. Likelihood score maximized using correct abundances

A key property of our framework is that the correct copy numbers (abundances) and assemblies maximizes our LAP score. To illustrate this property, we simulated a metagenomic community and calculated the LAP of the reference genomes with a combination of abundances. The simulated community consisted of *Bacillus cereus* and *Actinomyces odontolyticus* at a ratio of 4:7. We generated 200bp reads at 20x coverage of the metagenome (80x of *B. cereus* and 140x of *A odontolyticus*). We calculated the LAP scores of the error-free reference genomes for all combinations of abundances (ranging from 1 copy to 13 copies) for each bacteria.

We expect the highest LAP scores for the assemblies that contain the correct abundance ratio (4:7). As seen in Figure 1, our LAP score is able to accurately reflect the varying organism abundance ratios present in the sample. The closer the LAP scores to the correct abundance ratio, the higher the LAP scores with the true abundance ratios yielding the highest LAP scores in both communities.

B. Likelihood scores correlate with reference-based metrics

With real metagenomic samples, it is difficult to make evaluations given the lack of high quality references. Using purely simulated data has the issue of not accurately capturing the error and bias introduced by sequencing technology. Thus, to evaluate our LAP score, we use two 'mock' communities (Even and Staggered) provided by the Human Microbiome Project (HMP) consortium [2], [3]. These communities were created using specific DNA sequences from organisms with known reference genomes (consisting of over 20 bacterial genomes and a few eukaryotes) and abundances. We calculated the LAP score on assemblies produced by MetAMOS [14] using several assemblers: SOAPdenovo [16], Metavelvet [5], Velvet [17], and Meta-IDBA [4]. The additional *de novo* and reference-based metrics for the assemblies were taken from [14]. These metrics include: total number of contigs/scaffolds in the assembly (#ctgs), total amount of sequence (in Mbp) that can be aligned to the reference genomes (Aln), the size of the largest contig c such that the sum of all contigs larger than c is more than 10 Mbp (similar to the commonly used N50 size) (Sz @ 10Mbp), and average number of errors per Mbp (Err/Mbp). Additional referenced-based metrics not included in Table I can be found in [14].

In the mock Even dataset, the *de novo* LAP score agrees with the referenced-based metrics (Table I). SOAPdenovo has the greatest LAP score and total amount of sequence that can be aligned to a reference genome, while also having the lowest amount of misassemblies (including chimeric) and errors per Mbp. It is important to note that if user selected an assembly based on the best contiguity at 10Mbp, they would select the MetaVelvet assembly, which contains double the error rate per Mbp as the SOAPdenovo assembly while aligning 2Mbp less to the references.

Since the abundances of each organism in the mock Even dataset are fairly similar, the mock Staggered abundance distribution creates a more realistic scenario encountered in metagenomic environments. Here, the Meta-IDBA assembly

Table I
COMPARISON OF ASSEMBLY STATISTICS FOR HMP MOCK EVEN AND MOCK STAGGERED DATASETS.[1]

Data	Assembler	LAP	#ctgs	Aln	Sz @ 10Mbp	Err/Mbp	Aln reads
Even	SOAPdenovo	**-27.031**	63,107	**51**	28,208	**5.8**	**85.75%**
Even	Velvet	-28.537	12,830	41	42,269	8.7	83.30%
Even	MetaVelvet	-27.102	22,772	49	**62,138**	12.7	85.65%
Even	Meta-IDBA	-31.166	22,032	47	26,141	11	81.81%
Stag	SOAPdenovo	-60.161	44,928	**28**	5,672	**8.3**	69.78%
Stag	Velvet	-60.711	21,050	**28**	6,060	21.5	67.26%
Stag	MetaVelvet	-60.442	20,551	**28**	6,685	20.1	67.72%
Stag	Meta-IDBA	**-58.851**	4,559	18	**13,150**	10.2	**70.67%**

has the greatest LAP score, but aligns roughly a third less sequences to the reference genomes than SOAPdenovo. The Meta-IDBA assembly contains approximately a tenth of the amount of contigs (4,559 vs. 44,928) as SOAPdenovo. The SOAPdenovo assembly contains a greater percentage of contigs at a very low abundance. 75% of SOAPdenovo's contigs have an abundance of less than 10 compared to 0.95% in Meta-IDBA. On large contigs Meta-IDBA performs better than SOAPdenovo and has a lower error rate (see Figure 4 in [14]). However, Meta-IDBA assembles a smaller fraction of the low-abundance genomes than SOAPdenovo, leading fewer sequences to align. The LAP score penalizes misassemblies within abundant contigs in the SOAPdenovo results.

C. Tuning assembly parameters for MetAMOS

Assemblathon1 [8] has shown that assembly experts can often get drastically different assemblies using the same assemblers, highlighting the difficulty of choosing the *right* parameters for a given assembler. Our metagenomic LAP framework comes packaged with the MetAMOS pipeline, allowing users the option to run MetAMOS with different assemblers and automatically select the assembly with the highest LAP score. This step occurs without any prior knowledge from the user.

We showcase the ease of use of the automated assembler selection within MetAMOS using the *Carsonella ruddii* (156Kbp) dataset packaged with MetAMOS (Table II). Errors were found using DNADIFF [18]. The newbler assembly produced one contig containing the complete *C. ruddii* genome. The SOAPdenovo assembly produced a severely fragmented assembly with the most number of errors. The MetaVelvet and Velvet produced identical assemblies, containing 3 contigs of sizes 92Kbp, 65Kbp, and 1.7Kbp, but contained an additional 158bp compared to the *C. ruddii* genome. Upon closer inspection, there were overlaps between the contigs ranging from 38bp to 73bp. This is not surprising given MetaVelvet's and Velvet's de bruijn graph-based approach could not resolve repetitive regions between the contigs. Newbler, on the other hand, contained only a single insertion error. The LAP score of the Newbler assembly was greater due to more reads being able to align across the regions that were broken apart in the MetaVelvet and Velvet assemblies. Additionally, the Newbler assembly did not contain the duplicated sequence found in the other assemblies. MetAMOS was able to select the most likely assembly without requiring any additional input from the user.

[1]Numbers in bold represent the best value for the specific dataset.

Table II
SELF-TUNING METAMOS USING *C. ruddii* TEST DATASET.

Assembler	Contigs	LAP	N50 (Kbp)	Errors
newbler	**1**	**-13.064**	**156**	1
SOAPdenovo	23	-14.238	9	**3**
Velvet	3	-13.157	92	**0**
MetaVelvet	3	-13.157	92	**0**

IV. DISCUSSION

Results from GAGE [7] and Assemblathon [8], [9] have shown that the specific characteristics of the data being assembled has a great impact on the performance of the assembler. This problem is magnified in metagenomic assembly. By integrating our extended LAP framework into MetAMOS, we have allowed researchers to accurately and effortlessly run and evaluate assemblies without any prior knowledge on evaluating metagenomic assembly quality.

It should be noted that in some cases it may not be tractable to run the complete collection of assemblers with MetAMOS. In such cases, we should first employ heuristics (such as [19]) to aid in selecting potential assemblers (and parameters) to run. For the assembler selection process, we can use the LAP framework's sampling procedure in combination with calculating read probabilities in parallel to reduce runtime.

Our goal was to provide a global measure of how good a metagenomic assembly may be, not to detect assembly errors. Other likelihood-based frameworks, such as ALE, use frequencies of certain sequences to aid in detection of possible chimeric contigs. We are able to apply similar modifications to our LAP framework to find regions of possible misassembly. In addition, we plan to extend our framework to give a more detailed breakdown of the LAP scores of segments assembled using the same subset of reads across different assemblies. The goal would be to take high-scoring assembled segments from individual assemblies to recreate an assembly with overall greater likelihood. This approach will be of great benefit to the field of metagenomic assembly since assemblers are often designed with different constraints and goals in mind, e.g., low memory footprint, assembling high/low coverage organisms, or tolerating population polymorphisms. For example, on the mock Staggered dataset, Meta-IDBA best assembled the most abundant genomes while SOAPdenovo had a better representation of the low abundance organisms. Providing a systematic way of combining assembler approaches using our LAP score will produce better assemblies for downstream analyses.

V. Conclusion

In this paper we have described an extension to our *de novo* assembly evaluation framework (LAP) for accurately comparing metagenomic assemblies. In addition, we have integrated our framework into the metagenomic assembly pipeline MetAMOS, showing that any user is able to reproduce quality evaluations of metagenomic assemblies with relative ease.

Acknowledgment

The authors would like to thank the members of the Pop lab for valuable discussions on all aspects of our work.

This work was supported in part by the NIH, grant R01-AI-100947 to MP, and the NSF, grant IIS-1117247 to MP.

The contributions of SK and TJT were funded under Agreement No. HSHQDC-07-C-00020 awarded by the Department of Homeland Security (DHS) for the management and operation of the National Biodefense Analysis and Countermeasures Center (NBACC), a Federally Funded Research and Development Center. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security. In no event shall the DHS, NBACC, or Battelle National Biodefense Institute (BNBI) have any responsibility or liability for any use, misuse, inability to use, or reliance upon the information contained herein. The Department of Homeland Security does not endorse any products or commercial services mentioned in this publication.

References

[1] D. B. Rusch, A. L. Halpern, G. Sutton, K. B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J. A. Eisen, J. M. Hoffman, K. Remington *et al.*, "The sorcerer ii global ocean sampling expedition: northwest atlantic through eastern tropical pacific," *PLoS biology*, vol. 5, no. 3, p. e77, 2007.

[2] M. Mitreva *et al.*, "Structure, function and diversity of the healthy human microbiome," *Nature*, vol. 486, pp. 207–214, 2012.

[3] B. A. Methé, K. E. Nelson, M. Pop, H. H. Creasy, M. G. Giglio, C. Huttenhower, D. Gevers, J. F. Petrosino, S. Abubucker, J. H. Badger *et al.*, "A framework for human microbiome research," *Nature*, vol. 486, no. 7402, pp. 215–221, 2012.

[4] Y. Peng, H. C. Leung, S.-M. Yiu, and F. Y. Chin, "Meta-idba: a de novo assembler for metagenomic data," *Bioinformatics*, vol. 27, no. 13, pp. i94–i101, 2011.

[5] T. Namiki, T. Hachiya, H. Tanaka, and Y. Sakakibara, "Metavelvet: an extension of velvet assembler to de novo metagenome assembly from short sequence reads," *Nucleic acids research*, vol. 40, no. 20, pp. e155–e155, 2012.

[6] S. Boisvert, F. Raymond, É. Godzaridis, F. Laviolette, J. Corbeil *et al.*, "Ray meta: scalable de novo metagenome assembly and profiling," *Genome biology*, vol. 13, no. 12, p. R122, 2012.

[7] S. Salzberg, A. Phillippy, A. Zimin, D. Puiu, T. Magoc, S. Koren, T. Treangen, M. Schatz, A. Delcher, M. Roberts *et al.*, "Gage: A critical evaluation of genome assemblies and assembly algorithms," *Genome Research*, 2011.

[8] D. Earl, K. Bradnam, J. John, A. Darling, D. Lin, J. Fass, H. Yu, V. Buffalo, D. Zerbino, M. Diekhans *et al.*, "Assemblathon 1: A competitive assessment of de novo short read assembly methods," *Genome research*, vol. 21, no. 12, pp. 2224–2241, 2011.

[9] K. R. Bradnam, J. N. Fass, A. Alexandrov, P. Baranay, M. Bechner, İ. Birol, S. Boisvert10, J. A. Chapman, G. Chapuis, R. Chikhi *et al.*, "Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species," *arXiv preprint arXiv:1301.5406*, 2013.

[10] S. Clark, R. Egan, P. I. Frazier, and Z. Wang, "Ale: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies," *Bioinformatics*, 2013.

[11] A. Rahman and L. Pachter, "Cgal: computing genome assembly likelihoods," *Genome Biology*, vol. 14, no. 1, p. R8, 2013.

[12] M. Ghodsi, H. C.M., I. Astrovskaya, H. Lin, S. D., K. S., and M. Pop, "De novo genome assembly evaluation," In press.

[13] G. W. Tyson, J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar, and J. F. Banfield, "Community structure and metabolism through reconstruction of microbial genomes from the environment," *Nature*, vol. 428, no. 6978, pp. 37–43, 2004.

[14] T. J. Treangen, S. Koren, D. D. Sommer, B. Liu, I. Astrovskaya, B. Ondov, A. E. Darling, A. M. Phillippy, and M. Pop, "Metamos: a modular and open source metagenomic assembly and analysis pipeline," *Genome biology*, vol. 14, no. 1, p. R2, 2013.

[15] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with bowtie 2," *Nature methods*, vol. 9, no. 4, pp. 357–359, 2012.

[16] R. Li, H. Zhu, J. Ruan, W. Qian, X. Fang, Z. Shi, Y. Li, S. Li, G. Shan, and e. a. Kristiansen, K, "De novo assembly of human genomes with massively parallel short read sequencing," *Bioinformatics*, vol. 20(2), pp. 265–272, 2010.

[17] D. R. Zerbino and E. Birney, "Velvet: algorithms for de novo short read assembly using de bruijn graphs," *Genome Res.*, vol. 18, pp. 821–829, 2008.

[18] A. Phillippy, M. Schatz, and M. Pop, "Genome assembly forensics: finding the elusive mis-assembly," *Genome Biology*, vol. 9, no. 3, p. R55, 2008.

[19] R. Chikhi and P. Medvedev, "Informed and automated k-mer size selection for genome assembly." *Bioinformatics (Oxford, England)*, 2013.