Identifying Inter-Genomic Repeats Using Fast, Approximate Methods for Betweenness Centrality

Jay Ghurye
University of Maryland - College Park
jayg@cs.umd.edu

Christopher M. Hill
University of Washington School of Medicine
chrismh@uw.edu

Mihai Pop
University of Maryland - College Park
mpop@umiacs.umd.edu

1. INTRODUCTION

Genomic repeats are the most important challenge in genomic assembly even for isolate genomes. When reads are shorter than the repeats it can be shown that the number of genome reconstructions consistent with the read data grows exponentially with the number of repeats [4]. In the case of isolate genomes, long read technologies have largely addressed this challenge, at least for bacteria where the majority of genomic repeats fall within the range of achievable read lengths [5]. In metagenomics, however, the problem is compounded by the fact that microbial mixtures often include multiple closely-related genomes differing in just a few locations. The genomic segments shared by closely related organisms – inter-genomic repeats – are substantially larger than intra-genomic repeats and cannot be fully resolved even if long read data were available. Instead, the best hope is to identify and flag these repeats in order to avoid mis-assemblies that incorrectly span across genomes. In metagenomics, however, the problem is exacerbated by uneven coverage across contigs and by the fact that both inter- and intra-genomic repeats exist in the data. An alternate approach based on the social network concept of betweenness centrality was introduced in Bambus 2 [6]. In Bambus 2, for example, repeat finding based on exact betweenness centrality in a typical stool sample requires days of computation. To overcome this limitation, we explore the use of parallel and approximate betweenness centrality algorithms for repeat detection in metagenomic assembly graphs, and the effect of the level of approximation on the efficiency and accuracy of these algorithms.

2. BACKGROUND

In network analysis, metrics of centrality are used to identify the most important nodes within a graph. In this work, we use betweenness centrality. The betweenness centrality of a particular node is equal to the number of shortest paths from all nodes to all others that pass through that node. Brandes[2] proposed an exact algorithm for computing betweenness centrality of all the nodes based on single source shortest path approach. However, this approach is not suited for large metagenomic data. Several approaches for approximate betweenness centrality have been proposed. Bader and Pich[1] provide an approximation algorithm by choosing a subset of k starting nodes called pivots and apply the Brandes algorithm to just the chosen nodes. This algorithm was shown to overestimate the centrality of some unimportant nodes which are close to the pivots. Geisberger et al. [3] solve this issue by changing the scheme for aggregating betweenness contributions so that nodes close to the pivots are not unduly profited. A different approximation strategy was proposed by Riondato et al. [7] based on randomized sampling of shortest paths (rather than nodes), approach which offers probabilistic guarantees on the quality of approximation. This algorithm guarantees that all approximate values of betweenness for all vertices are within an additive factor $\epsilon \in (0, 1)$ from the real values with probability at least $1 - \delta$.

3. METHODS

Data

We evaluate the efficacy of detecting inter-genomic repeats using the approximate betweenness centrality methods on simulated metagenomic assembly graphs. We start with a collection of genomes from which we construct idealized De Bruijn graphs (assuming perfect coverage and no sequencing errors). Next we modify the De Bruijn graph generated from the collection of genomes to identify repeats. We simulated metagenomic assembly graphs consisting of two, five, and ten bacterial genomes. For the five and ten metagenome samples, random bacterial genomes were chosen from the data set provided by [8]. A k-mer size of 55 was chosen for the initial creation of the De Bruijn graph.

Cutoff Criteria For Repeats Identification

Centrality algorithms provide numeric estimates for the centrality of each node, values from which we can infer whether a graph node represents a repeat. In Bambus 2 [6] repeats were defined as nodes with a centrality score larger than three standard deviations from the mean centrality of the nodes in the graph. We use the same definition here and also explore the use of an alternate statistic independent of underlying distribution – the *interquartile range(IQR)*. Let Q_1,Q_2, and Q_3 denote the lower, middle and upper quartiles respectively. IQR can be defined as the difference between the upper and lower quartile. So, $IQR = Q_3 - Q_1$. To identify repeats we simply mark all the nodes with centrality values larger than $Q_3 + 1.5 * IQR$.

4. RESULTS

Across all datasets, the efficiency of the approach is correlated with the level of approximation for both node sampling and path sampling approaches (Figs. 1 and 2). Accuracy is inversely proportional with the level of approximation. The effectiveness of these methods is also higher in simpler communities. In the 2-genome community, all inter-genomic

repeats were found by sampling as few as ten nodes, an order of magnitude less time than calculating the betweenness centrality using the full graph. For the 5-genome graph, sampling as few as ten nodes yields a sensitivity of 86% and full sensitivity is achieved with just 500 nodes, or approximately one tenth of the entire size of the assembly graph.

The 10-genome graph is substantially more complex yielding much lower sensitivity for both node sampling and path sampling approaches. The use of the inter-quartile range as a decision criterion improves the sensitivity from about 6% to 35% in the 10-genome graph at the cost of a reduction in specificity from 99.3% to 90.9%. Also, note that the path sampling approach is marginally more efficient than the node sampling procedure - the worst runtime (at an error setting of 1%) is roughly equivalent to sampling about 1000 nodes.

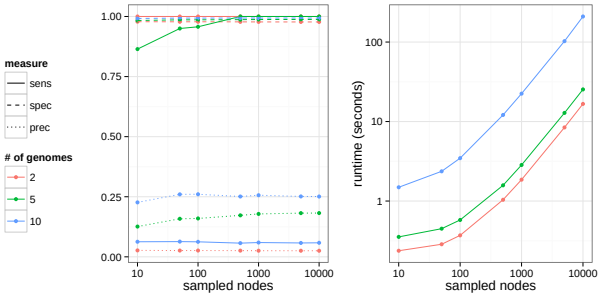

Figure 1: Statistical measures of repeat detection quality for the simulated metagenomic assembly graphs. Betweenness centrality was approximated by randomly sampling nodes in the graph.

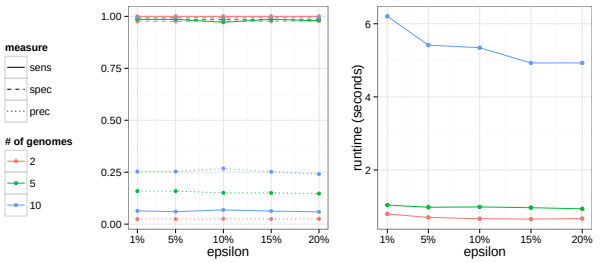

Figure 2: Statistical measures of repeat detection quality for the simulated metagenomic assembly graphs. Betweenness centrality was approximated within epsilon of their true value.

5. CONCLUSION AND FUTURE WORK

Betweenness centrality is a proven method for finding inter-genomic repeats in metagenomic assemblies. The size of typical metagenomic assembly graphs often make it infeasible to calculate the exact betweenness centrality of each node, as demonstrated by the slow runtime of the MarkRepeats procedure from Bambus 2. Here we have demonstrated the effectiveness of approximate measures of centrality able to identify repeats in a fraction of the time previously required. We explored two alternative approximation strategies, based on node sampling and path sampling, respectively. The runtime of node-sampling approaches can be more effectively estimated as it increases roughly linearly in the size of the sample selected. Conversely, path sampling approaches can guarantee the level of approximation but runtime cannot be easily estimated.

Despite promising results, our work has also revealed limitations of the approximate approaches - in the more complex communities the sensitivity of detection dropped significantly, though it was partly rescued by the use of a decision cut-off based on inter-quartile ranges. We are currently exploring this phenomenon and whether it will have a significant impact on metagenomic assembly. It is likely that many of the repeats missed by the approximate procedure are small and local in nature and could be resolved through other means. It is also possible that other approaches for outlier detection would be more effective in restoring the sensitivity of detection. We currently plan to incorporate such algorithms in the Bambus scaffolding software in order to improve its efficiency, and also plan to further develop the repeat detection algorithms in order to improve their sensitivity in complex graphs.

6. REFERENCES

[1] D. A. Bader, S. Kintali, K. Madduri, and M. Mihail. Approximating betweenness centrality. In *Algorithms and Models for the Web-Graph*, pages 124–137. Springer, 2007.

[2] U. Brandes. A faster algorithm for betweenness centrality*. *Journal of Mathematical Sociology*, 25(2):163–177, 2001.

[3] R. Geisberger, P. Sanders, and D. Schultes. Better approximation of betweenness centrality. In *ALENEX*, pages 90–100. SIAM, 2008.

[4] C. Kingsford, M. C. Schatz, and M. Pop. Assembly complexity of prokaryotic genomes using short reads. *BMC bioinformatics*, 11(1):21, 2010.

[5] S. Koren and A. M. Phillippy. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current opinion in microbiology*, 23:110–120, 2015.

[6] S. Koren, T. J. Treangen, and M. Pop. Bambus 2: scaffolding metagenomes. *Bioinformatics*, 27(21):2964–2971, 2011.

[7] M. Riondato and E. M. Kornaropoulos. Fast approximation of betweenness centrality through sampling. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 413–422. ACM, 2014.

[8] M. Shakya, C. Quince, J. H. Campbell, Z. K. Yang, C. W. Schadt, and M. Podar. Comparative metagenomic and rrna microbial diversity characterization using archaeal and bacterial synthetic communities. *Environmental microbiology*, 15(6):1882–1899, 2013.