MICHIGAN STATE
UNIVERSITY

December 16th, 2011

Dear Editors,

We would like the editors to consider our manuscript entitled “Scaling metagenome sequence assembly with probabilistic de Bruijn graphs” for publication in the Proceedings of the National Academy.

In this manuscript, we address the pressing bioinformatic problem of assembling sequence data from extremely complex microbial populations such as those in soil. Advances in DNA sequencing have made it possible to deeply sample such populations, but bioinformatic approaches have not kept up. To tackle this problem, we introduce and analyze a novel de Bruijn assembly graph representation that lets us scale the assembly of one particular soil sample by a factor of over 20 over current approaches. In support of our conclusions, we provide a theoretical approach, software implementing the approach, and the results of executing this software.

The predominant method of scaling metagenome assembly for complex metagenomes is to discard low-abundance data, an approximate technique that biases against rare species. This has been used in three recent publications:

Mackelprang et al., Nature. 2011 Nov 6;480(7377):368-71.
Qin et al., Nature. 2010 Mar 4;464(7285):59-65.
Hess et al., Science. 2011 Jan 28;331(6016):463-7.

However, these environments are all relatively simple compared to the soil environments now being sequenced, which are estimated to contain over a million species per gram of soil (Gans et al., Science. 2005 Aug 26;309(5739):1387-90.). To completely sample a single gram would require over 5 terabases (5e12) of sequencing data, as compared to the less than 300 gigabases (3e11) incompletely analyzed in the above publications. It is clear that current assembly approaches simply cannot handle this volume of data.

Our results suggest that a computational technique known colloquially as “divide and conquer” can be applied to the problem of scaling metagenome assembly, and we demonstrate a general approach, a data structure, and an implementation of both on simulated data and a real data set. Most importantly, our approach relies on no approximations or heuristics to subdivide the data set; we demonstrate identical results before and after our approach is applied to a data set. We believe that the partitioning approach

COLLEGE OF ENGINEERING

Department of Computer Science and Engineering

Michigan State University
3115 Engineering Building
East Lansing, Michigan
48824-1226

(517) 353-3148
FAX: (517) 432-1061

developed in this paper will become a standard method for achieving sequence assemblies of large, complex metagenomes. Moreover, we know of no other similarly effective approaches to scaling metagenome assembly that do not discard data.

More broadly, this computational technique and the associated data structure should apply to transcriptome assembly, and there are potentially many other uses of a lower-memory de Bruijn graph data structure that remain to be explored. Current methods in bioinformatics are generally struggling to deal with the vast amount of data now readily available from next-generation sequencing machines, and our data structure addresses this challenge.

The paper uses concepts from graph theory to analyze a new use of a computer science data structure for bioinformatic use in tackling a biological problem. For reviewers, we suggest:

Mihai Pop, Associate Professor, U. Maryland, mpop@umiacs.umd.edu
Haixu Tang, Associate Professor, Indiana University, hatang@indiana.edu
Lior Pachter, Professor, UC Berkeley, lpachter@berkeley.edu
Folker Meyer, Associate Directory, Institute for Genomics and Systems Biology, Argonne National Lab, folker@anl.gov
John Wooley, Professor, UC San Diego, jwooley@ucsd.edu.

Sincerely,

C. Titus Brown (corresponding author)
Assistant Professor
Computer Science and Engineering /
Microbiology and Molecular Genetics
Michigan State University