Analysis of 16S Genomic Data using Graphical Databases

Conference Paper · August 2017

DOI: 10.1145/3107411.3108208

4 authors, including:

Olivia Ahern
University of Rhode Island
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Rebecca J. Stevick
University of Rhode Island
4 PUBLICATIONS **1** CITATION
SEE PROFILE

Analysis of 16S Genomic Data using Graphical Databases

Olivia M. Ahern
University of Rhode Island
Graduate School of Oceanography
Narragansett, Rhode Island 02882
olivia_ahern@uri.edu

Rebecca J. Stevick
University of Rhode Island
Graduate School of Oceanography
Narragansett, Rhode Island 02882
rstevick@my.uri.edu

Li Yuan
University of Rhode Island
Department of Computer Science and Statistics
Kingston, Rhode Island 02881
li_yuan@my.uri.edu

Noah M. Daniels*
University of Rhode Island
Department of Computer Science and Statistics
Kingston, Rhode Island 02881
noah_daniels@uri.edu

1 INTRODUCTION

Since the Human Genome Project was completed in 2003, many data scientists have developed algorithms in order to store and query high volumes of genomic data. The most common data storage techniques employed in these algorithms are flat files or relational databases. While sophisticated indexing techniques can accelerate queries [5], an alternative is to store biological sequence data directly in a way that supports efficient queries. Here we introduce a new algorithm that aims to compress the redundant information and improve the performance of query speed with the help of graphical databases, which have been commercial available since the mid-late 2000s. A graphical database stores information using nodes and relationships (edges).

Our approach is to identify subsequences that are common among many sequences, and to store these as "common nodes" in the graphical database. This is accomplished for sequencing data as follows: split the whole sequence into k-mers: if a given k-mer is common to enough sequences, then it is labeled as a *common segment*; if a k-mer is unique (or common to too few sequences), then it is labeled as a *single segment*. Thus, common nodes and single nodes are formed from common segments and single segments, respectively. These two kinds of nodes are connected by edges in the graphical database, allowing each original sequences to be reconstructed by following edges in the graph.

This graphical database model allows for fast taxonomic queries of 16S rDNA. When queried, the database can first attempt to find common nodes that match the query sequence, and subsequently follow edges to single nodes to refine the search. This approach is analogous to that of "compressive genomics" [2, 4] except that the compression is implicit in the graphical database storage model.

Beyond simple sequence queries, this graphical database representation also supports variability analysis [1], which identifies highly variable vs. conserved regions of 16S sequence. Regions of low variability correspond to common nodes, while regions of high variability correspond to a variety of paths through single nodes. Figure 1 illustrates common and single nodes, and a corresponding plot of variability.

Benchmarking of sequence search indicates that query time in graphical databases is significantly faster than in flat files or relational databases. Implementation of graphical databases in genomic data analysis will allow for accelerated search, and may lend itself to other forms of efficient analysis, such as tetramer frequency analysis, which is useful in metagenomic binning [3].

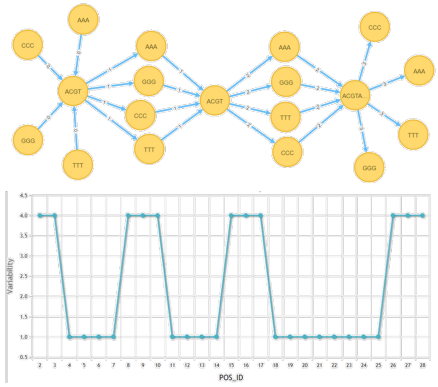

Figure 1: Example of common and unique nodes in a 16S rDNA dataset. The variability plot depicts the number of nodes found at each position in sequence. Common nodes correspond to low variability.

REFERENCES

[1] Josselin Bodilis, Sandrine Nsigue-Meilo, Ludovic Besaury, and Laurent Quillet. 2012. Variable copy number, intra-genomic heterogeneities and lateral transfers of the 16S rRNA gene in Pseudomonas. *PloS one* 7, 4 (2012), e35647.
[2] Noah M Daniels, Andrew Gallant, Jian Peng, Lenore J Cowen, Michael Baym, and Bonnie Berger. 2013. Compressive genomics for protein databases. *Bioinformatics* 29, 13 (2013), i283–i290.
[3] Hsin-Hung Lin and Yu-Chieh Liao. 2016. Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. *Scientific reports* 6 (2016).
[4] Po-Ru Loh, Michael Baym, and Bonnie Berger. 2012. Compressive genomics. *Nature biotechnology* 30, 7 (2012), 627–630.
[5] Jared T Simpson and Richard Durbin. 2010. Efficient construction of an assembly string graph using the FM-index. *Bioinformatics* 26, 12 (2010), i367–i373.

*to whom correspondence should be addressed

ACM-BCB'17, August 20–23, 2017, Boston, MA, USA.
 ISBN 978-1-4503-4722-8/17/08.
DOI: http://dx.doi.org/10.1145/3107411.3108208