Improving Metagenomic Assemblies Through Data Partitioning: A GC Content Approach

Fábio Miranda[1(✉)], Cassio Batista[1], Artur Silva[2,3], Jefferson Morais[1], Nelson Neto[1], and Rommel Ramos[1,2,3]

[1] Computer Science Graduate Program, Federal University of Pará, Belém, Brazil
{fabiomm,cassiotb,jmorais,nelsonneto,rommelramos}@ufpa.br
[2] Institute of Biological Sciences, Federal University of Pará, Belém, Brazil
asilva@ufpa.br
[3] Center of Genomics and Systems Biology, Federal University of Pará, Belém, Brazil

Abstract. Assembling metagenomic data sequenced by NGS platforms poses significant computational challenges, especially due to large volumes of data, sequencing errors, and variations in size, complexity, diversity and abundance of organisms present in a given metagenome. To overcome these problems, this work proposes an open-source, bioinformatic tool called GCSplit, which partitions metagenomic sequences into subsets using a computationally inexpensive metric: the GC content. Experiments performed on real data show that preprocessing short reads with GCSplit prior to assembly reduces memory consumption and generates higher quality results, such as an increase in the size of the largest contig and N50 metric, while both the L50 value and the total number of contigs produced in the assembly were reduced. GCSplit is available at https://github.com/mirand863/gcsplit.

AQ1

Keywords: DNA sequencing · Metagenomics · Data partitioning
Bioinformatic tools · Metagenomic data preprocessing

1 Introduction

Metagenomics consists in determining the collective DNA of microorganisms that coexist as communities in a variety of environments, such as soil, sea and even the human body [1–3]. In a sense, the field of metagenomics transcends the traditional study of genes and genomes, because it allows scientists to investigate all the organisms present in a certain community, thus allowing the possibility to infer the consequences of the presence or absence of certain microbes. For example, sequencing the gastrointestinal microbiota enables the understanding of the role played by microbial organisms in the human health [4].

Nevertheless, second generation sequencing technologies—which belong to the Next Generation Sequencing (NGS), and are still the most widespread technology on the market—are unable to completely sequence the individual genome

I. Rojas and F. Ortuño (Eds.): IWBBIO 2018, LNBI 10813, pp. 1–11, 2018.
https://doi.org/10.1007/978-3-319-78723-7_36