Journal of Computational Science xxx (2013) xxx–xxx

Contents lists available at ScienceDirect

Journal of Computational Science

journal homepage: www.elsevier.com/locate/jocs

An approach to a metagenomic data processing workflow

Milko Krachunov[a,*], Dimitar Vassilev[b]

[a] Faculty of Mathematics and Informatics, University of Sofia "St. Kliment Ohridski", 5 James Bourchier Blvd., 1164 Sofia, Bulgaria
[b] Bioinformatics Group, AgroBio Institute, 8 Dragan Tsankov Blvd., 1164 Sofia, Bulgaria

ARTICLE INFO

Article history:
Received 15 January 2013
Received in revised form 16 June 2013
Accepted 13 August 2013
Available online xxx

Keywords:
Metagenomics
Error detection
NGS data analysis workflow

ABSTRACT

Metagenomics is a rapidly growing field, which has been greatly driven by the ongoing advancements in high-throughput sequencing technologies. As a result, both the data preparation and the subsequent *in silico* experiments pose unsolved technical and theoretical challenges, as there are not any well-established approaches, and new expertise and software are constantly emerging.

Our project main focus is the creation and evaluation of a novel error detection and correction approach to be used inside a metagenomic processing workflow. The approach, together with an indirect validation technique and the already obtained empirical results, are described in detail in this paper. To aid the development and testing, we are also building a workflow execution system to run our experiments that is designed to be extensible beyond the scope of error detection which will be released as a free/open-source software package.

1. Introduction

1.1. Problems in metagenomics

Metagenomics deals with the mixed genetic data found in samples collected from heterogeneous biological environments, ranging from soil to the insides of various macro-organisms. These microbial communities are still largely unexplored, presenting the researchers with samples containing a large number of organisms from a great variety of microbial species, a large portion of which are presently unknown.

Comparative analysis of these microbial communities is crucial for studies that explore issues ranging from human health [1] to bacterial and viral evolution [2]. They can have an impact on our understanding of the past of our biosphere as well as dealing with potential future threats – as the most rapidly mutating agents, microbes can provide a lot of insight on evolution, and are also a critical factor in unexpected disease outbreaks.

A researcher in the field of metagenomics has to deal with a variety of challenges [3,4]. As a new field there are yet no well-established methods to approach it, and they often have to face unsolved technical or methodological problems. The datasets are large and heterogeneous, most of the microbial species comprising them are not sequenced elsewhere, and with their rate of mutation it is unclear if the present means of cataloguing genomes can be a feasible approach to simplify this task.

Due to the nature of the data obtained – which presently lacks any inherent reference points or a standard for validation – every study involves the computational challenges associated with high-throughput *de novo* sequencing, which are further exacerbated by the need to deal with a larger degree of uncertainty, significantly larger amount of data and the need to adapt the data processing to every particular experiment, often multiple times.

At present, all researchers have to deal with deficiencies in the data quality and limited capabilities of the software tools and processing methods. Their work involves time consuming processing of huge datasets and a great deal of uncertainty about the correctness of the input data as well as the results.

1.2. Our project and goal

Initially, our project began as an attempt to reduce the impact of errors on the quality of the metagenomic studies by proposing a new error detection approach and comparing it with other approaches on metagenomic data. Soon it became clear, however, that obtaining a pristine metagenomic test data set, that can be used to give a definite confirmation of the advantage of one error detection method over another, could prove to be very difficult if not impossible because of the difficulty in taking the same sample again. To deal with this, we had to come up with roundabout approaches to indirectly estimate the number of false positives and false negatives that an error detection procedure suffers from. These approaches, however, are neither as reliable nor as easy as a direct measurement of the quality of a real dataset. As a result, they

* Corresponding author. Tel.: +359 885988001.
E-mail addresses: elsevier@milko.3mhz.net (M. Krachunov), jim6329@gmail.com (D. Vassilev).

1877-7503/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jocs.2013.08.003

```
TCTCTATGCGCCATTGTAGCACGTGTGTAGCC...
TCTCTATGCGCCATGTAGCACGTGTGTAGCC...
```

Fig. 1. An excerpt from the input datasets.

are dependent on the execution of a big number of computational experiments on a very large number of datasets.

These experiments constitute a processing workflow that executes multiple genomic software packages in which both the parameters and the procedure need to be varied. We came to the conclusion that building a tool for managing, running and distributing the genomic toolchain would greatly reduce the amount of manual work required to run any metagenomic experiment.

Thus our initial goal of building and validating error detection was extended to the larger project of developing a library for executing configurable genomic workflows capable of interfacing with arbitrary external tools.

2. Material and methods

2.1. The input data

16S RNA is very attractive for metagenomic analysis, because it is highly conserved and thus largely similar across a great deal of species, while at the same time it contains hypervariable regions that are incredibly helpful for identifying species, individual organisms and finding their evolutionary relationships [5].

The sample datasets for our experiments contain short reads between 300 and 500 bases in length, divided in sets of tens of thousands of sequences – between 20000 and 50000 after filtering them by length (≥300, ≤500 bp) and quality (throwing out ambiguous bases). All our sample datasets were sequenced using the 454 platform by Roche. It is very suitable for metagenomic experiments because it produces short reads of sufficient length.

2.2. Data preparation using sequence alignment

One of the crucial steps in a typical metagenomic workflow is the sequence alignment and any processing heavily relies on one's ability to do fast multiple sequence alignments of acceptable quality. If we look at any sample excerpt from our datasets like the one in Fig. 1 we can easily notice that the sequences are displaced because of missing or extra bases. This makes it impossible to perform any meaningful column-wise analysis unless such displacements are accounted for by the sequence alignment.

A high-quality alignment is particularly important for the execution of the error detection approach proposed in the next section which relies on column-wise comparison across multiple species. If alignment of the denoised data is desired, a modification of the error detection method extended to perform correction can be used on a preliminary alignment before the real alignment is executed.

Unfortunately for metagenomics the datasets are much larger and far more varied than those found in regular genomics, and the approaches for alignment used in *de novo* sequencing, resequencing and sequence searching are no longer suitable [6–8].

Finding the globally optimal alignment for n sequences is an NP-complete problem. For any considerably sized dataset like the ones found in metagenomics finding this optimum is a practical impossibility. Furthermore, inexact methods are more likely to produce bad results for heterogeneous data. This is in contrast with the highly similar data that is found in genomic studies where the sequences are usually limited to a single species, or where there are means for obtaining representative sequences that input data can be aligned against, which allows one to align fast without sacrificing the quality.

In our experiments, we ran several alignment software packages intended for large datasets. We discovered that when we ran them with the stricter parameters intended for higher quality results they could not process our data in a reasonable time (the execution time was in the range of days), but when we ran them in a less accurate mode they did not produce acceptable results.

To remedy this we used a surprisingly simple and straightforward approach. We performed a quick clustering of the dataset using the CD-HIT-454 software [9]. We aligned each cluster with a software solution and settings for a high-quality alignment, in particular we used MAFFT [10] and MUSCLE [11]. Then, we aligned the clusters against each other and combined them in a manner similar to the one used in multiple sequence alignment using a guide tree.

Counter-intuitively, the alignment took significantly less time and was significantly superior in quality to the alignment that we got when we ran MAFFT or MUSCLE directly. While evidently the alignment of the metagenomic datasets is feasible, we did not find a straightforward solution and we had to improvise despite the fact alignment is a very basic component of the metagenomic processing.

Such makeshift solutions are not always obvious and can differ greatly in quality depending on how they are constructed, and as such can be greatly facilitated by software for building and launching preconfigured workflows. Such software would also allow for a quicker comparison between the various options as this would no longer need to be done by hand. One of our major goals is not simply to build a metagenomic workflow or pipeline that performs multiple sequence alignment, but to extend it as to allow easier experimentation by allowing arbitrary combinations of software packages to perform this task.

2.3. Improving read quality by error detection and correction

One significant obstacle in metagenomic studies is the uncertainty about the data correctness. Sequencing equipment produces a great deal of errors that can be intermixed with meaningful differences with biological origin such as mutations and meaningless errors with biological origin such as errors during amplification, all of which initially occur randomly.

The mutations are an important subject of evolutionary studies and can provide invaluable insight on the development and propagation of microbial species. Unlike the other two kinds of errors, mutations are most often found at an evolutionary dead-end that kills the organism, which makes the surviving ones peculiar in that they are an object of interest with the information about the species they carry, while at the same time they provide an opportunity for distinguishing them from actual errors.

A common approach to tell them apart is to use their frequency of appearance, which is not always reliable. It is common practice to throw out any reads suspected to have errors in them, but this can reduce the size of the dataset by an order of magnitude, while most of the discarded information was correct.

Improving the means for detecting and correcting those errors, as well as proposing ways to utilise the information present in those often discarded sequences, is one thing that can lead to a significant improvement in all metagenomic studies.

2.3.1. The naïve approach

The most obvious way to spot errors is simply look for data that occurs rarely. This can be done by counting the frequency of occurrence of each base in each column. The bases that appear less often than a threshold that was established beforehand are considered errors.

The assumption behind this approach is that while mutations happen at a slower rate then errors, their numbers are multiplied by inheritance, as the surviving ones will span through multiple