A Fast Microorganism Detection Algorithm based on High-Throughput Sequencing Data

LI Jiangyu，MAO Yiqing，WANG Xiaolei，LIU Yang，ZHAO Dongsheng
Institute of Health Service and Medical Information, Academy of Military Medical Sciences,
Beijing 100850, China

I. INTRODUCTION

High-throughput sequencing speed is faster than before[1]. By sequencing, it is possible to research microorganisms which are difficult to cultivate. Many diseases are caused by microorganisms such as bacteria and virus. For example, 12% of human cancers are caused by viruses[2]. Therefore, the detection of microorganisms is particularly important. Some microorganism detection algorithms are released recently, such as PathSeq[3], RINS[4], VirusSeq[5], CaPSID[6], et al.

RINS uses the priori knowledge algorithm to reduce the data processed, so it is much faster than PathSeq. But if the assumption is proved to be wrong when using RINS, then users need to reassume, analyze and verify the data. RINS is weak in finding the microorganisms with no reference genomes. While PathSeq depends on Amazon's EC2 cloud computing platform [7] and S3 cloud storage platform[8], the use of PathSeq is influenced by the network, and PathSeq takes more time to process. Microorganism detection algorithm can synthesize the two ideas to achieve better performance.

II. ALGORITHM FOR MICROORGANISM DETECTION

The detection algorithm is designed as Fig. 1.

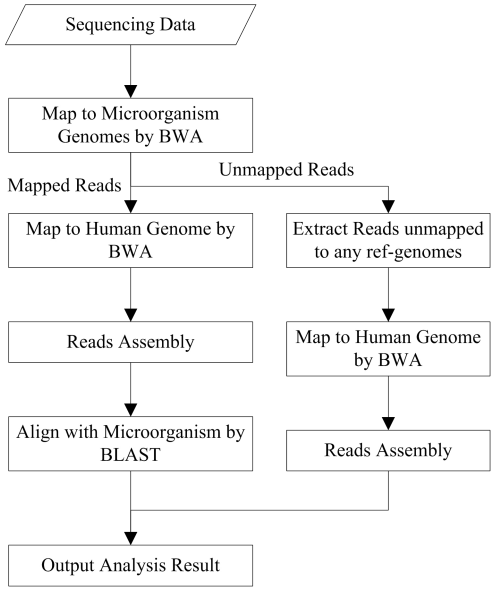

Figure 1. The pipeline of microorganism detection algorithm

Short read aligner is used to map the sequencing data in the above steps. By comparison, BWA[9] can get more mapped reads than Bowtie[10], while the processing speed of the two software is very close. Our algorithm uses BWA to improve accuracy. Velvet[12] is used in the algorithm as short reads assembler, which is based on De Bruijn graph[11]. Assembled contigs are long reads. Our algorithm uses BLAST[13] to perform the alignment.

III. RESULT

Hg19[16] is used as reference human genome. Microbial reference genomes are downloaded from NCBI[17] from viruses, bacteria and fungi categories (Update to Sep., 2012)

Simulated sequencing datasets: simulated sequencing data are generated by GemSIM [18], described as follows.

(1) Known microbial sequencing data: It means there is reference genome for the microorganism in the sample. GemSIM uses the human genome hg19 and the bacteria genome Helicobacter_pylori_Sat464_uid159467 as templates to generate simulated data. Generated reads are paired-end sequencing data, and read length is 101bp.

(2) Unknown microbial sequencing data: It means there is no reference genome for the microorganism in the sample. GemSIM uses the human genome hg19 and a simulated mutated virus sequence as templates to generate simulated data. Reads generated are paired-end, and the read length is 101bp. Virus Hepatitis B virus isolate HK2100 is used as the original sequence to generate mutated virus sequence. The mutation rate is 80%.

Real sequencing data: the dataset is SRR073726, which is human CA-HPV-10 prostate cancer sequencing data. The sequencing reads are paired-end sequencing data, with read length 40bp. The serum samples contain Human Papilloma Virus subtype 18 (HPV18)

Results for the known microbial sequencing data are described in Table 1

TABLE I. RESULT FOR KNOWN MICROBIAL SEQUENCING DATA

Data Group	Runtime (s)	Contigs Number	Length>10 kbp	Longest contig	Avg. Length
Top 20%	758	2134	0	3293	1337
Top 40%	963	310	46	53168	15451

Top 60%	1216	143	40	130802	35677
Top 80%	1474	127	37	130802	39174
100%	1738	128	37	130802	39174

These contigs are aligned to microbial genomes by BLASTN. According to the map result, microbial genome gi|384893616|ref|NC_017359.1 | Helicobacter pylori Sat464 chromosome gets the highest match score with these contigs.

Results for the unknown microbial sequencing data are shown in Table 2.

TABLE II. RESULT FOR UNKNOWN MICROBIAL SEQUENCING DATA

Data Group	Runtime(s)	Contigs(bp)	Mapped Bases(bp)
Top 20%	533	1636, 1431	1636, 1431
Top 40%	782	2113, 960	2113, 960
Top 60%	1019	3074	3074
Top 80%	1257	2114, 961	2114, 960
100%	1522	3136	3076

Results for the real sequencing data are shown in Table 3.

TABLE III. RESULT FOR REAL SEQUENCING DATA

Data Group	Runtime(s)	Contigs(bp_	Mapped Bases(bp)
Top 20%	118	600, 182	583, 126
Top 40%	126	590, 180	58p, 175
Top 60%	223	592, 186	583, 181
Top 80%	356	592, 186	58p, 181
100%	441	592, 186	583, 181

IV. DISCUSSION

The algorithm described in the paper first reference RINS algorithm to extract the microbial sequencing data in the data processing, which can reduce the amount of sequencing data need to be processed. Meanwhile, in the algorithm, if the analysis results of the extracted data of the virus sequence fail to pass the verification, then in turn change the reference genomes to bacteria and fungi to extract the bacterial and fungus sequencing data for analysis and verification. The purpose is to avoid analyzing partial data when replacing the reference genomes. This algorithm can also analyze the remaining reads after the extraction of microbial genomes sequencing data, thereby reducing the probability of missing unknown microorganism.

When using the sequencing-by-side method to deal with the three test datasets, our detection algorithm gets short contigs for the first obtained sequencing data, with the increase of sequencing data, contigs obtained by the algorithm gradually become longer. When the amount of sequencing data is 60% of the total amount of data, the detection algorithm gets almost the same result The results of sequencing-by-side analysis show that the method can obtain preliminary test results with the initial sequencing data of the sample. Along with the increase of the sequencing data, the results become more reliable.

[1] llumina Website. An Introduction to Next-Generation Sequencing Technology. http : //www.illumina.com/Documents/products/Illumina_Sequencing_Introduction.pdf.

[2] Hausen Z. “The Search for Infectious Causes of Human Cancers: Where and Why”. Virology, 2009, 392:1-10.

[3] Kostic AD, Ojesina AI, Pedamallu CS, et al. “PathSeq : software to identify or discover microbes by deep sequencing of human tissue”. Nature Biotechnology, 2011, 29(5): 393-396.

[4] Bhaduri A, Qu K, Lee CS, et al. “Rapid Identification of Nonhuman Sequences in High Throughput Sequencing Data Sets”. Bioinformatics, 2012, 28(8): 1174-1175.

[5] Chen YX, Yao H, Thompson EJ, et al. “VirusSeq: Software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue”. Bioinformatics, 2013, 29(2): 266-267.

[6] Borozan I, Wilson S, Blanchette P, et al. “CaPSID: A bioinformatics platform for computational pathogen sequence identification in human genomes and transcriptomes”. BMC Bioinformatics, 2012, 13: 206-217.

[7] Introduction for Amazon EC2 cloud computing platform. http://aws.amazon.com/cn/ec2/.

[8] Introduction for Amazon S3 cloud storage[EB/OL]. http://aws.amazon.com/cn/s3/.

[9] Li H, Durbin R. “Fast and accurate short read alignment with Burrows-Wheeler Transform”. Bioinformatics, 2009, 25: 1754-60.

[10] Langmead B, Trapnell C, Pop M, Salzberg SL. “Ultrafast and memory-efficient aligment of short DNA sequencing to the human genome”. Genome Biology, 2009, 10(3): R25.

[11] Rodriguez N, Hackenberg M, Aransay AM. “Bioinformatics for High Throughput Sequencing”. Springer Science+Business Media, 2012: 90-103.

[12] Zerbino D, Birney E. “Velvet: algorithms for de novo short read assembly using de Bruijn graphs”. Genome Research, 2008, 18: 821-829.

[13] Altschul SF, Gish W, Miller W, et al. “Basic local alignment search tool”. Journal of Molecular Biology, 1990, 215 (3):403−410.

[14] MiSeq Personal Sequencer. http://www.illumina.com/systems/miseq.ilmn.

[15] BaseSpace. http://www.illumina.com/software/basespace.ilmn?sciid=2011019IBN2&utm_campaign=illumina.com_homepage_banner_small_internal&utm_medium=banner&utm_source=illumina.com.

[16] Hg19. ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19.

[17] NCBI. www.ncbi.nlm.nih.gov/.

[18] McElroy KE, Luciani F, Thomas T. “GemSIM: general, error-model based simulator of next-generation sequencing data”. BMC Genomics, 2012, 13: 74.