A Fast Short Read Alignment Algorithm Using Histogram-based Features

Qiu Chen, Koji Kotani[*], Feifei Lee, and Tadahiro Ohmi

New Industry Creation Hatchery Center, Tohoku University
[*] *Department of Electronics, Graduate School of Engineering, Tohoku University*
Aza-Aoba 6-6-10, Aramaki, Aoba-ku, Sendai 980-8579, JAPAN

e-mail: qiu@fff.niche.tohoku.ac.jp

***Abstract*—Current new generation of DNA sequencers have had the ability to generate billions of short reads rapidly and inexpensively. How to solve fast and robust short read alignment problem become one of the most important challenges in bioinformatics research area. The current solutions for short-read alignment have limitations that alignment algorithms such as MAQ and Bowtie have few capabilities to align reads with insertions or deletions. In this paper, we propose an efficient hierarchical alignment algorithm to reduce it. For a given short read, first, a fast histogram search method is used to scan the reference sequence. Most of locations in reference sequence with low similarity will be excluded for latter searching. The Smith-Waterman alignment algorithm is then applied to each remainder location to search for exact matching. Experimental results show the proposed method combining histogram information and Smith-Waterman algorithm is a faster and accurate algorithm for short read alignment.**

Keywords-Short read; Alignment; Fast search; Smith-Waterman; Histogram-based feature

I. INTRODUCTION

The decipherment of 3-billion-base human genome sequence was finally completed by the international cooperation in April 2003 [1][2]. Since this achievement of human genome project, researchers around the world are now having a very keen competition on clarification of the structure and performance analysis of the protein, genes and protein networks, and new gene sequences are clarified every day. The enormous quantity of data has been accumulated in the database like GenBank [3], EMBL [4], and DDBJ [5], etc. Moreover, the volume of data of Genome Database still increases in exponential [6].

Current new generation of DNA (Deoxyribonucleic Acid) sequencers have had the ability to generate billions of short reads rapidly and inexpensively. The Illumina/Solexa sequencing technology typically generates 50-200 million 32-100 bp reads on a single run of the machine [13], which is transforming genomic science. These new machines are quickly becoming the technology of choice for whole-genome sequencing and for a variety of sequencing-based assays, including gene expression, DNA-protein interaction, human resequencing and RNA splicing studies [7].

In resequencing, a reference genome is already available for the species and one is interested in comparing short reads obtained from the genome of one or more donors (individual members of the species) to the reference genome. Therefore, the first step in any kind of analysis is the mapping of short reads to a reference genome.

How to map large amount of short reads to a reference sequence (e.g., the human genome) has become a challenging topic to the existing sequence alignment programs. A lot of new alignment algorithms have been developed to meet the requirement of efficient and accurate short read mapping.

Current available main algorithms for short read alignment include Bowtie [9], SOAP [10], SOAP2 [11], MAQ [12], BWA [13], mrFAST [14], mrsFAST [15], Novoalign [16] and SHRiMP [17], etc.

There are 4 types of the DNA nucleotides, namely, A (adenine), C (cytosine), G (guanine) and T (thymine), which are utilized to encode DNA. Due to sequencing errors and/or genetic variations, many reads map to the reference sequence approximately but not exactly, and therefore, to map a read to the reference sequence, read mapping programs should allow a certain number of mismatches between the read and a candidate location.

But current solutions for short-read alignment have limitations while implementing in an actual alignment application. SOAP [10], Novoalign [16], etc. can be easily parallelized with multi-threading, but large memory are usually necessary for building an index for the human genome [13]. SOAP2 [11], MAQ [12], Bowtie [9] and BWA [13] have few capabilities to align reads with insertions or deletions. If number of mismatches increases, it will take more time to align billions of reads to a large reference, and the accuracy will be reduced.

In this paper, we propose an efficient hierarchical alignment algorithm using Histogram-based Features and Smith-Waterman algorithm (HF-SW) that can tolerate moderate mismatches. For a given short read, first, a fast histogram search method is used to scan the reference sequence. Most of locations in reference sequence with low similarity will be excluded for latter searching. The Smith-Waterman alignment algorithm [18] is then applied to each remainder location to search for the exact matching. The effects will be demonstrated by using simulated data as well as real data.

This paper is organized as follows. In Section II, we will first introduce the proposed alignment algorithm using histogram-based features for short read in detail. Experimental results using both simulated data and real data

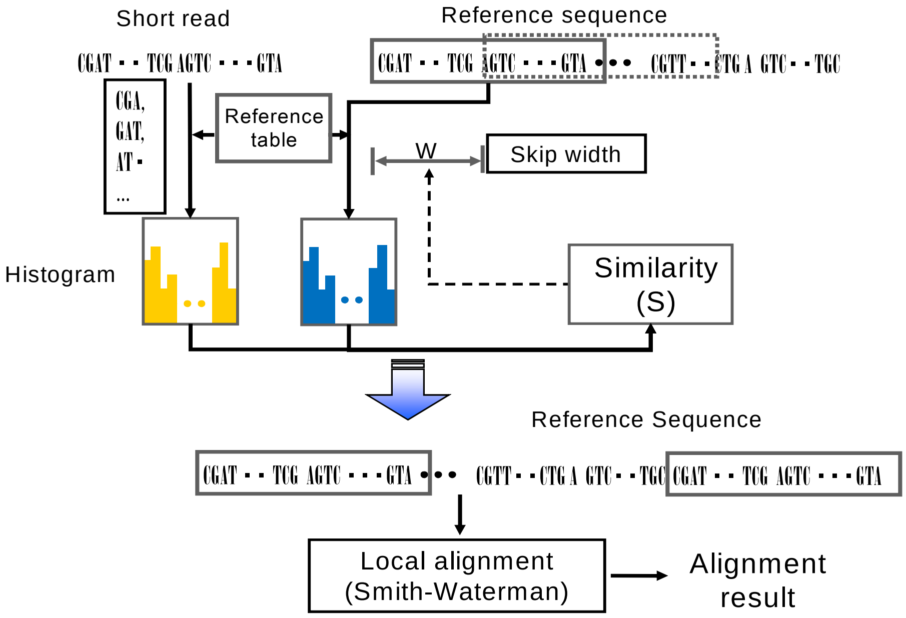
Short read
CGAT • • TCG AGTC • • • GTA
CGA,
GAT,
AT •
...
Reference table
Reference sequence
CGAT • • TCG AGTC • • • GTA • • CGTT • • CTG A GTC • • TGC
W
Skip width
Histogram
Similarity (S)
Reference Sequence
CGAT • • TCG AGTC • • • GTA • • CGTT • • CTG A GTC • • TGC CGAT • • TCG AGTC • • • GTA
Local alignment (Smith-Waterman)
Alignment result

Figure 1. Processing steps of proposed method.

will be discussed in Section III. Finally, conclusions are given in Section IV.

II. Proposed Method

In this paper, we present a new short read alignment algorithm for short reads mapping in a large size of reference sequence. Histogram-based features of a given short read are firstly used to compare with the reference sequence and similarity scores would be obtained. Only the locations whose similarities exceeded a given threshold are then aligned using exhaustive Smith-Waterman dynamic programming algorithm [18].

Table I. Reference table.

CCC	CCT	CCG	CCA	CTC	CTT	CTG	CTA
0	1	2	3	4	5	6	7
CGC	CGT	CGG	CGA	CAC	CAT	CAG	CAA
8	9	10	11	12	13	14	15
TCC	TCT	TCG	TCA	TTC	TTT	TTG	TTA
16	17	18	19	20	21	22	23
TGC	TGT	TGG	TGA	TAC	TAT	TAG	TAA
24	25	26	27	28	29	30	31
GCC	GCT	GCG	GCA	GTC	GTT	GTG	GTA
32	33	34	35	36	37	38	39
GGC	GGT	GGG	GGA	GAC	GAT	GAG	GAA
40	41	42	43	44	45	46	47
ACC	ACT	ACG	ACA	ATC	ATT	ATG	ATA
48	49	50	51	52	53	54	55
AGC	AGT	AGG	AGA	AAC	AAT	AAG	AAA
56	57	58	59	60	61	62	63

Figure 1 shows the processing steps of our proposed method. When an unknown short read is input, it will be divided into small sequence, for instance, ACT and CGG, etc. A small sequence can be considered as a three dimensional vector. This processing overlaps over the entire short read. After that, the histogram feature is calculated. There are only 4 types of DNA bases, so the number of combination of 3-dimensional vector is 64. A reference table with the size of 64 is shown in Table I, by which the index number of the 3-dimensional vector is very easy and fast to be determined. The number of vectors with same index number in each separate partial sequence is counted and feature vector histogram is easily generated, and it is used as histogram feature of the short read.

In the mapping stage, the windows are applied to both the short read and the reference sequence. Corresponding histogram-based features of the short read and the partial sequence of the reference sequence in the window are generated as described above. The similarity between these histograms is then calculated. If the similarity exceeded a threshold value given previously, the location candidate will be detected and located. Otherwise, the window on the reference sequence will be skipped to the next position determined by the similarity in current position and the threshold value. In the last step, the window on the reference sequence is shifted forward and the mapping proceeds.

Here, histogram intersection is used as the similarity measure [20], and is defined as formula (1).

$$\begin{aligned} S_{SR} &= S(H_S, H_R) \\ &= \frac{1}{N}\sum_{l=1}^{L}\min(h_{Sl}, h_{Rl}) \end{aligned} \qquad (1)$$

where h_{Sl}, h_{Rl} are the numbers of feature vectors contained in the *l-th* bin of the histograms for the short read and the partial reference sequence, respectively, L is the number of histogram bins, and N is the total number of feature vectors contained in the histogram. The skip width w is shown by formula (2).

$$w = \begin{cases} floor(N(\theta - S_{SR})) + 1 & (S_{SR} < \theta) \\ 1 & otherwise \end{cases} \qquad (2)$$

where *floor(x)* means the greatest integral value less than *x*, and θ is a given threshold.

When reference sequence scanning is finished, location candidates whose similarities exceeded a given threshold are selected. The Smith-Waterman alignment algorithm [18] is then applied to each remainder location to search for the exact matching.

III. Experiments and Discussions

To evaluate our proposed short read alignment algorithm using Histogram-based Features and Smith-Waterman algorithm (HF-SW), we compared its performance with one of the main short read alignment algorithm named Burrows-Wheeler Alignment tool (BWA) which achieves better performance than other main alignment algorithms such as Bowtie [7], SOAP2 [9], and MAQ [10].

BWA is a new read alignment package that is based on backward search with Burrows–Wheeler Transform (BWT), to efficiently align short sequencing reads against a large reference sequence such as the human genome, allowing mismatches and gaps. For short read alignment against the human reference genome, BWA is an order of magnitude faster than MAQ while achieving similar alignment accuracy [13].

We performed all of the experiments on a conventional PC@3.2GHz (12G memory). The algorithm was implemented in ANSI C.

A. Evaluation on simulated data

We simulated reads from the human genome using the wgsim program that is included in the SAMtools package [19] and ran the both programs to map the reads back to the human genome. Because the exact coordinate of each read, we are able to calculate the alignment error rate.

Table II shows that HF-SW achieved similar alignment accuracy with error rate of 0.117% as BWA at short read length of 70. HF-SW is more accurate than BWA when short read length is longer than 125.

The mapping time spending of HF-SW algorithm at 70 bp is about 854 seconds, which is similar with BWA. As shown in Table II, HF-SW algorithm can finish mapping only 937 seconds at 125 bp, which is about 2.2 times faster than BWA algorithm.

Table II. Comparison between BWA and proposed method using emulated data.

Algorithm	Time(s)	Err(%)
BWA-32	569	0.3
HF-SW-32	746	0.53
BWA-70	1093	0.12
HF-SW-70	854	0.117
BWA-125	2104	0.05
HF-SW-125	937	0.044

Table III. Comparison between BWA and proposed method using real data.

Algorithm	Time(h)	Conf(%)
BWA-51	3.2	88.9
HF-SW-51	3.15	88.7

B. Evaluation on real data

To evaluate the performance on real data, we downloaded about 12.2 million pairs of 51 bp reads from European Read Archive (AC:ERR000589). These reads were produced by Illumina for NA12750, a male included in the 1000 Genomes Project. Reads were mapped to the human genome NCBI build 36.

As shown in Table III, HF-SW confidently mapped 88.7% of all reads in 3.15 hours, which achieved similar performance compared with BWA algorithm.

IV. Conclusions

In this paper, we proposed a novel short read alignment algorithm combining histogram features and Smith-Waterman dynamic programming algorithms. Experimental results using emulated data as well as real data show proposed alignment algorithm will give more robust resulting and the proposed method is more efficient compared with conventional algorithms for short read alignment.

REFERENCES

[1] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, et al., “The sequence of the human genome,” Science, vol. 291, no. 5507, pp. 1304 -1351, 2001.

[2] F. S. Collins, M. Morgan, and A. Patrinos, “The human genome project: lessons from Large-Scale Biology,” Science, vol. 300, no. 5617, pp. 286-290, 2003.

[3] GenBank, ftp://ftp.ncbi.nih.gov/genbank/.

[4] EMBL, http://www.embl.org/

[5] DDBJ, http://www.ddbj.nig.ac.jp/

[6] http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html.

[7] C. Trapnell and S. L. Salzberg, “How to map billions of short reads onto genomes,” Nature Biotechnology, vol. 27, 2009, pp. 455-457.

[8] Q. Chen, K. Kotani, F. Lee, and T. Ohmi, “A Fast Retrieval of DNA Sequences Using Histogram Information,” 2009 Int’l Conf. on Future Information Technology and Management Engineering (FITME 2009), pp. 529-532, Sanya, China, Dec., 2009.

[9] B. Langmead, C. Trapnell, M. Pop and S. L. Salzberg, “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome,” Genome Biol., vol. 10, 2009, R25.

[10] R. Li, Y. Li, K. Kristiansen, and J. Wang, “SOAP: short oligonucleotide alignment program,” Bioinformatics, vol. 24, 2008, pp. 713-714.

[11] R. Li, C. Yu, Y. Li, T. W. Lam, S. M. Yiu, K. Kristiansen, and J. Wang, “SOAP2: an improved ultrafast tool for short read alignment,” Bioinformatics, vol. 25, 2009, pp. 1966-1967.

[12] H. Li, J. Ruan, and R. Durbin, “Mapping short DNA sequencing reads and calling variants using mapping quality scores,” Genome Res., vol. 18, 2008, pp. 1851-1858.

[13] H. Li and R. Durbin, “Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics,” vol. 25, no. 14, 2009, pp. 1754–1760.

[14] C. Alkan, J. M. Kidd, T. Marques-Bonet, G. Aksay, F. Antonacci, F. Hormozdiari, et al., “Personalized copy number and segmental duplication maps using next-generation sequencing,” Nature Genetics, vol. 41, 2009, pp. 1061-1067.

[15] F. Hach, F. Hormozdiari, C. Alkan, F. Hormozdiari, I. Birol, E. E. Eichler, and S C. Sahinal, “mrsFAST: a cache-oblivious algorithm for short-read mapping,” Nature Methods, vol.7, 2010, pp. 576-577.

[16] Novocraft, http://www.novocraft.com/.

[17] K.R. Rasmussen, J. Stoye, and E. W. Myers, “Efficient q-gram filters for finding all e-matches over a given length,” Lecture Notes in Computer Science, Springer, vol. 3500, 2005, pp. 189-203.

[18] T. F. Smith and M. S.Waterman, “Identification of common molecular subsequences”, Journal of Molecular Biology, vol. 47, pp. 195-197, 1981.

[19] SAMtools, http://samtools.sourceforge.net.

[20] V.V. Vinod and H. Murase, “Focused color intersection with efficient searching for object extraction”, Pattern Recognition, vol. 30, no.10, 1997, pp. 1787-1797.

[21] 1000 Genomes Project, http://www.1000genomes.org.