Optics and Laser Technology 108 (2018) 124–135

Contents lists available at ScienceDirect

Optics and Laser Technology

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

Full length article

Speeding up DNA sequence alignment by optical correlator

Fereshte Mozafari, Hossein Babashah, Somayyeh Koohi *, Zahra Kavehvash

Sharif University of Technology, Iran

ARTICLE INFO

Article history:
Received 23 October 2017
Received in revised form 3 May 2018
Accepted 11 June 2018

Keywords:
Optical computing
Metamaterial
Fourier optics
Bioinformatics
DNA sequence alignment

ABSTRACT

In electronic computers, extensive amount of computations required for searching biological sequences in big databases leads to vast amount of energy consumption for electrical processing and cooling. On the other hand, optical processing is much faster than electrical counterpart, due to its parallel processing capability, at a fraction of energy consumption level and cost. In this regard, this paper proposes a correlation-based optical algorithm using metamaterial, taking advantages of optical parallel processing, to efficiently locate the edits as a means of DNA sequence comparison. Specifically, the proposed algorithm partitions the read DNA sequence into multiple overlapping intervals, referred to as windows, and then, extracts the peaks resulted from their cross-correlation with the reference sequence in parallel. Finally, to locate the edits, a simple algorithm utilizing number and location of the peaks is introduced to analyze the correlation outputs obtained from window-based DNA sequence comparison. As a novel implementation approach, we adopt multiple metamaterial-based optical correlators to optically implement the proposed parallel architecture, named as Window-based Optical Correlator (WOC). This wave-based computing architecture fully controls wave transmission and phase using dielectric and plasmonic materials. Design limitations and challenges of the proposed architecture are also discussed in details. The simulation results, comparing WOC with the well-known BLAST algorithm, demonstrate superior speed-up up to 60%, as well as, high accuracy even at the presence of large number of edits. Also, WOC method considerably reduces power consumption as a result of implementing metamaterial-based optical computing structure.

1. Introduction

Comparing characters of DNA sequences against a database of reference (i.e., consensus) is defined as DNA sequence alignment. High throughput sequencing (HTS) technologies were introduced in 2006 [1], and the latest iterations of HTS technologies have an ability to read the genome of a human during three days by the cost of $1000 [2].

The Illumina/Solexa sequencing technology [3] typically produces 50–200 million 32–100 base-pairs (bps) reads by only one running [4]. After generating the short reads, they must be mapped (i.e., align) to a known reference genome. The mapping process is computationally very expensive, since the reference genome is very large (e.g., the human genome has 3.2 G bps). So, mapping large volume of short reads to a genome as large as a human genome is a serious challenge for the existing sequence alignment programs.

A mapper is the name of one software that performs the mapping. The mapper has to search a very large reference genome database to map millions of short reads. On the other hand, each short read may possess edits defined as the difference between bps from read and reference fragment, representing either deletions, insertions, or mismatches, which necessitates expensive approximate searching. To simplify searching a large database, such as a human genome, various studies have developed several algorithms for mappers [5,6].

Considering electrical implementation, several mappers have been developed over the past few years, which can be classified based on their mapping algorithms into two categories; (1) hash table based or seed-and-extend mappers similar to the well-known BLAST [7] method, such as mrFAST/mrsFAST [8,9], MAQ [10], SHRiMP [11], Hobbes [12], drFAST [13] and RazerS [14]; and (2) suffix-array or genome compression based mappers that use the Burrows-Wheeler Transform (BWT) such as BWA [4], Bowtie [15], and SOAP2 [16]. Each category of mapping algorithms has its own strengths and weaknesses. For performance evaluation of different mappers in finding best alignments, three general metrics are considered: speed of alignment algorithm,

* Corresponding author at: Sharif University of Technology, Tehran, Tehran 11155-9517, Iran.
E-mail address: koohi@sharif.edu (S. Koohi).

https://doi.org/10.1016/j.optlastec.2018.06.027
0030-3992/

accuracy of the algorithm in aligning reads including multiple edits, and finally, comprehensiveness in searching for all aligning locations across the reference genome. The hash table-based mappers compared to the suffix-array based mappers are much slower, more accurate, more comprehensive, and more robust to sequence edits and genomic diversity. For these reasons, hash table based mappers for comparing the different species genomes, are typically preferred, such as mapping reads from a gorilla genome by the human reference genome [3]. On the other hand, suffix-array based mappers, with the BWT optimization, benefits high mapping speed, up to 30-fold faster than hash table based mappers, while by increasing edit distance between the read and the reference fragment, both their mapping accuracy and comprehensiveness reduce. In single nucleotide polymorphism (SNP) discovery studies, where accuracy is less important, suffix-array based mappers are preferred because of their speed [5]. All the aforementioned electronic based mapper algorithms suffer from upper electronic speed limit of transistors. Therefore, it would be suitable to design an architecture that utilizes parallel processing resulting in speed increment.

Many parallel processing architectures have been proposed including FPGA [17], GPU [2] and Optics [18–20]. Performance of electronic-based computing, especially in the case of big data processing, is usually limited by high power consumption and inevitably low speed of serial processing [21]. Whereas, optical computing benefits from parallel processing inherited in optics and low power loss such as methods that presented in [22]. Accordingly, among all of the aforementioned approaches, optics is the only solution that can utilize parallelism with no limit and also benefits low power consumption.

Although optics offers high speed processing as a result of parallel processing, it suffers from some challenges due to its implementation considerations. Specifically, since any arbitrary algorithm cannot be efficiently implemented in optics, any algorithm introduced for sequence alignment should meet its limitations and benefit from the nature of optical processing. Therefore, electronic algorithms and architectures cannot be roughly implemented in optics, making it necessary to present an appropriate algorithm for optical implementation. However, traditional optical components are usually bulky and have slow response [18]. Therefore, new optical components that benefit from small size and fast response, namely metamaterial based components, are of the great interest [23]. Metamaterial based components are artificially engineered structures that can manipulate the impinging light in a sub-wavelength regime and perform computation. In [18], the concept of computational metamaterial is introduced to perform optical mathematical operations, including spatial differentiation, integration, and convolution. In this approach, manipulating the impinging wave to the desired output is obtained as the wave propagates through the metamaterial structure in order to carry out optical computing process. Finally, this paper proposes a method which not only supports high accuracy and comprehensiveness, but also overcomes electronics speed limitation by using optical parallel processing in less power consumption as well.

So far, various approaches such as optical correlation [24] and Moiré matching techniques [25,26] have been proposed to perform DNA sequence alignment by the means of optical computing process. In correlation based methods [24,27–29], DNA sequence alignment is performed through an optical correlator to detect regions with high score of similarity (referred to as global alignment). However, detecting the exact location of edits (referred to as local alignment) which is very critical in the field of bioinformatics [7], is not addressed by optical correlators. On the other hand, through the Moiré based matching techniques [25,30], the authors have attempted to exploit edit detection based sequence alignment algorithms. Although the proposed algorithm offers fairly acceptable performance for detecting edit locations, its applicability to big data analysis is not addressed. To address above limitations, in this paper, we propose an optical DNA alignment algorithm capable of both global and local alignments, as well as big data analysis.

In this paper, computational metamaterials are used to perform optical DNA sequence alignment based on optical correlator. For increasing the accuracy of global alignment and detecting exact location of edits, the read sequence is divided into several overlapped parts which are called windows. Implementing correlator with metasurfaces (i.e. ultrathin metamaterial) is performed in the Fourier domain by applying a transfer function proportional to Fourier transform of the reference genome sequence to the impinging optical signal of query genome sequence Fourier transform [18]. The Fourier transformation is also carried out with an input GRIN lens while a metasurface is placed in its focal plane. The analysis for various edit location detection is also considered and the performance of our method is compared against the state of the art electronic based architectures [4,7]. The simulation results demonstrate that the proposed method is promising for DNA sequence alignment operations. In summary, this paper makes the following contributions:

- Proposes a high speed optical correlation algorithm capable of accurate global alignment and locating edits by considering different window sizes.
- Proposes a correlation-based algorithm that its implementation is applicable in optics.
- Provides an optical structure for the new algorithm.
- Gives a comparison of the proposed optical architecture with the well-known BLAST sequence alignment algorithm. The comparison addresses the run-time, hardware complexity, and accuracy of proposed method against those of BLAST mechanism.

The rest of this paper is structured as follows: Section 2 presents the proposed algorithm including details of alignment and detecting the exact locations of edits, as well as the proposed optical structure. Section 3 discusses the simulation results. Finally, the paper is concluded in Section 4.

2. Proposed parallel optical DNA sequence alignment method

For decades, the alignment methods of Dayhoff [31], Smith-Waterman [32], and Needleman-Wunsch [33] have been enhanced and refined. Although faster methods, such as FASTA [34] and BLAST [7] algorithms, exist, they generally impose trade-os between high speed on one hand, and decreased accuracy or requirement for preprocessing of sequence dataset on the other hand. To overcome the aforementioned limitations, this paper proposes a novel sequence alignment approach based on optical cross-correlation, as an ultra-fast process. Although, various electronic processing approaches have been proposed so far based on different algorithms, our proposed architecture is an optically-assisted system implementing an efficient matching algorithm in optical domain. As a key advantage of the proposed optical approach, it considerably accelerates DNA processing and increases throughput, compared to its electronic counterpart, taking advantage of parallelism inherent in optical processing [21]. Summarizing above discussion, this paper makes several noteworthy contributions as follows:

- Proposing a high speed optical correlator by adopting metamaterials.