УДК 811.111

PROGRESS IN GENOMIC RESEARCH WITH BWA: EFFICIENCY, APPLICATIONS IN BIOTECHNOLOGY (РОЗВИТОК ГЕНОМНИХ ДОСЛІДЖЕНЬ З BWA: ЕФЕКТИВНІСТЬ, ЗАСТОСУВАННЯ В БІОТЕХНОЛОГІЇ)

Ященко А.Д. – здобувач вищої освіти групи БТ 3/1
Науковий керівник: Саламатіна О.О., доцент кафедри іноземних мов МНАУ

У цій публікації досліджується метод вирівнювання Берроуза-Вілера (BWA), основний метод біоінформатики, і детально описується його сучасне використання перетворення Берроуза-Вілера та FM-індексу для ефективного і точного вирівнювання послідовностей ДНК з референтними геномами.

Ключові слова: *біоінформатика, BWA, вирівнювання послідовностей, ДНК.*

This publication explores the Burrows-Wheeler Aligner (BWA), a cornerstone bioinformatics method, detailing its sophisticated use of the Burrows-Wheeler Transform and FM-index for the efficient and accurate alignment of DNA sequences to reference genomes.

Key words: *bioinformatics, BWA, sequence alignment, DNA.*

The analysis of genomic data plays a pivotal role in contemporary biotechnology, driving advancements in fields ranging from personalized medicine and pharmacogenomics to agricultural improvements and microbial genomics. By decoding the complex information encoded in the DNA, researchers can uncover genetic variations linked to diseases, identify targets for new drugs, develop genetically modified organisms for enhanced agricultural productivity, and much more. This wealth of genomic data offers unprecedented opportunities for understanding life at a molecular level and tackling some of the most pressing health and environmental challenges.

However, the sheer volume and complexity of genomic data present significant computational challenges. The rapid increase in genomic data outpaces traditional data processing capabilities, necessitating efficient and scalable computational tools for data analysis. One of the primary tasks in genomic data analysis is sequence alignment, where newly sequenced DNA fragments (reads) are aligned to a reference genome to identify genetic variants and understand genomic structures.

Among the tools developed to address these challenges, the Burrows-Wheeler Aligner (BWA) stands out as a cornerstone in genomic research. BWA utilizes the Burrows-Wheeler Transform (BWT) and the FM-index, sophisticated algorithms that allow for efficient and rapid alignment of DNA sequences to a reference genome. This capability is crucial for processing the vast amounts of data generated by high-throughput sequencing technologies. BWA supports various types of alignments, including end-to-end (traditional alignment) and local (allowing for partial alignments), making it versatile for different genomic research needs [1].

BWA has significantly contributed to the field by enabling the handling of large-scale genomic datasets, thus facilitating the identification of genetic markers for diseases, understanding population genetics, and exploring the genetic basis of traits in agriculture. Its efficiency and accuracy in aligning sequences to reference genomes have made it an indispensable tool in genomics research. However, the continuous growth in data volume and the evolving complexity of genomic studies call for ongoing enhancements in algorithms like BWA and the development of new computational strategies to keep pace with the advancing frontiers of genomics.

Essence of the BWA Method

The core of BWA’s methodology is the Burrows-Wheeler Transform (BWT) coupled with the FM-index, a compressed full-text substring index. This combination allows BWA to perform efficient, memory-saving alignments of DNA sequences (reads) to a reference genome. The BWT reorganizes the reference genome in a way that groups similar character sequences together, making it more amenable to compression and faster to search. The FM-index then enables quick, efficient searches of the transformed data, allowing BWA to locate the positions where genomic reads could align with the reference genome [3].

Problems Addressed by BWA

High-Throughput Sequencing Data Volume: With the advent of high-throughput sequencing technologies, the volume of genomic data has exploded. BWA addresses the challenge of aligning billions of short DNA sequences (reads) to large reference genomes quickly and with limited computational resources.

Accuracy and Speed: BWA balances the trade-off between speed and accuracy. It achieves fast alignment times without significantly compromising the accuracy of the alignments, which is crucial for downstream analyses like variant calling and genotyping [3].

Scalability: The method is scalable, capable of handling genomic data as reference genomes grow in size and complexity, and as sequencing technologies continue to increase the volume of data generated.

Flexibility in Alignment Types: BWA supports both end-to-end and local alignments, catering to various research needs. End-to-end alignment is used when the entire read must align from one end to the other, useful for many variant discovery applications. Local alignment, which allows for partial alignments, is beneficial when reads may contain insertions, deletions, or errors towards the ends.

Complex Genomic Regions: Through its efficient indexing and search algorithms, BWA can effectively align reads to repetitive or low-complexity regions of the genome, which are often challenging for other aligners. This capability is crucial for comprehensive genomic analyses, including those areas that are rich in genetic diversity and underlie many genetic diseases.

One exciting area for further investigation involves the integration of BWA with machine learning algorithms. The accuracy and speed of BWA make it an ideal candidate for preprocessing data for machine learning models, which can predict genetic diseases, discover novel genetic associations, and understand gene function. By combining BWA’s alignment capabilities with the predictive power of machine learning, researchers can uncover deeper insights into genomic data, leading to advances in personalized medicine and targeted therapies [4].

Another potential research direction focuses on improving alignment algorithms for emerging sequencing technologies. As sequencing technologies evolve, producing longer reads with unique error profiles, adapting and optimizing BWA to accommodate these changes is crucial. Research could explore modifications to BWA's algorithms to enhance its efficiency and accuracy for new data types, ensuring that it remains a valuable tool for genomic analysis in the era of long-read sequencing and beyond [5].

The exploration of parallel computing and cloud-based solutions presents a further opportunity. Given the exponential growth of genomic data, leveraging cloud computing and parallel processing can significantly reduce the time and computational resources required for genomic analyses. Future research could develop more sophisticated BWA implementations that are optimized for cloud environments, facilitating large-scale genomic studies and collaborative research efforts across the globe.

Literature:

1. BURROWS. *University of south florida.* URL: https://wiki.rc.usf.edu/index.php/Burrows. (date of access: 11.03.2024).
2. Burrows-Wheeler Aligner. Burrows-Wheeler Aligner. URL: https://bio-bwa.sourceforge.net/ (date of access: 11.03.2024).
3. GitHub - lh3/bwa: Burrow-Wheeler Aligner for short-read alignment (see minimap2 for long-read alignment). GitHub. URL: https://github.com/lh3/bwa (date of access: 11.03.2024).
4. José M. A. BigBWA: approaching the Burrows–Wheeler aligner to Big Data technologies, Bioinformatics, Volume 31, Issue 24, December 2015, Pages 4003–4005
5. Taheri, M., Ansari, M.S., Magierowski, S. and Mahani, A. (2021), Hardware acceleration of the novel two dimensional Burrows-Wheeler Aligner algorithm with maximal exact matches seed extension kernel. IET Circuits Devices Syst, 15: 94-103.